DMSVIVA 2023

Proceedings of the 29th International DMS Conference on Visualization and Visual Languages

June 29 to July 3, 2023
Larkspur Landing South San Francisco Hotel, USA and KSIR Virtual Conference Center, USA

DOI: 10.18293/DMSVIVA2023

Proceedings preparation, editing and printing are sponsored by KSI Research Inc.

PROCEEDINGS

DMSVIVA2023

The 29th International DMS Conference on Visualization and Visual Languages

Sponsored by

KSI Research Inc. and Knowledge Systems Institute, USA

Technical Program

June 29 to July 3, 2023

Larkspur Landing South San Francisco Hotel, USA and KSIR Virtual Conference Center, USA

Organized by

KSI Research Inc. and Knowledge Systems Institute, USA

ISSN: 2326-3261 (print)
2326-3318 (online)
DOI: 10.18293/DMSVIVA2023

Additional copies can be ordered from:
KSI Research Inc.
156 Park Square Lane
Pittsburgh, PA 15238 USA
Tel: +1-412-606-5022
Fax: +1-847-679-3166
Email: dms@ksiresearch.org
Web: http://ksiresearch.org/ seke/dmsviva23.html

Proceedings preparation and editing are sponsored by KSI Research Inc. and Knowledge Systems Institute, USA.

DMSVIVA2023

The 29th International DMS Conference on Visualization and Visual Languages

June 29 to July 3, 2023

Larkspur Landing South San Francisco Hotel, USA and KSIR Virtual Conference Center, USA

Conference Organization

DMSVIVA2023 Conference Chair and Co-Chairs

Bernardo Breve, University of Salerno, Italy; Conference Co-chair
Jun Kong, North Dokota State University, USA; Conference Co-chair

DMSVIVA2023 Steering Committee Chair

Shi-Kuo Chang, University of Pittsburgh, USA; Steering Committee Chair

DMSVIVA2023 Steering Committee

Paolo Nesi, University of Florence, Italy; Steering Committee Member
Kia Ng, University of Leeds, UK; Steering Committee Member

DMSVIVA2023 Program Chair

Walter Balzano, University of Naples, Italy; Program Chair

DMSVIVA2023 Program Committee

Danilo Avola, University of Rome, Italy
Andrew Blake, University of Brighton, UK
Paolo Bottoni, Universita Sapienza, Italy
Paolo Buono, University of Bari, Italy
Loredana Caruccio, University of Salerno, Italy
Maiga Chang, Athabasca University, Canada
Yuan-Sun Chu, National Chung Cheng University, Taiwan
WilliamCheng-Chung Chu, Tunghai University, Taiwan

Stefano Cirillo, University of Salerno, Italy
Mauro Coccoli, University of Genova, Italy
Gennaro Costagliola, University of Salerno, Italy
Mattia DeRosa, University of Salerno, Italy
Domenico Desiato, University of Salerno, Italy
Vincenzo Deufemia, University of Salerno, Italy
Tiansi Dong, Bonn-Aachen International Center for Information Technology, Germany
Martin Erwig, Oregon State University, USA
Larbi Esmahi, Athabasca University, Canada
Edoardo Fadda, Politecnico di Torino, Italy
Filomena Ferrucci, University of Salerno, Italy
Andrew Fish, University of Brighton, UK
Manuel Fonseca, University of Lisbon, Portugal
Rita Francese, University of Salerno, Italy
Kaori Fujinami, Tokyo University of Agriculture and Technology, Japan
Angela Guercio, Kent State University, USA
Pedro Isaias, University of Queensland, Australia
Jonathan Kavalan, University of Florida, USA
Jun Kong, North Dokota State University, USA
Yau-Hwang Kuo, National Cheng Kung University, Taiwan
Robert Laurini, University of Lyon, France
Alan Liu, National Chung Cheng Univeristy, Taiwan
Weibin Liu, Beijing Jiao Tung University, China
Mark Minas, Universität der Bundeswehr München, Germany
Andrea Molinari, University of Trento, Trento, Italy
Eloe Nathan, Northwest Missouri State University, USA
Paolo Nesi, University of Florence, Italy
Max North, Southern Polytechnic State University, USA
Michela Paolucci, University of Florence, Italy
Antonio Piccinno, Univ. of Bari, Italy
Giovanni Pilato, Italian National Research Council, Italy
Giuseppe Polese, University of Salerno, Italy
Elvinia Riccobene, University of Milano, Italy
Peter Rodgers, University of Kent, UK
Domenico Santaniello, University of Salerno, Italy
Michael Wybrow, Monash University, Australia
Weiwei Xing, Beijing Jiao Tung University, China
Atsuo Yoshitaka, JAIST, Japan
Tomas Zeman, Czech Technical University, Czech Republic
Yang Zou, Hohai University, China

Publicity Chair and Co-Chair

Maiga Chang, Athabasca University, Canada; Publicity Co-Chair
Tiansi Dong, Bonn-Aachen International Center for Information Technology, Germany; Publicity Co-Chair

FOREWORD

On behalf of the Program Committee of the *29th International DMS Conference on Visualization and Visual Languages (DMSVIVA2023)*, I would like to welcome you. This conference aimed at bringing together experts in visualization, visual languages, distance education and distributed multimedia computing, providing a forum for productive discussions about these topics.

I would like to thank all the authors for their contributions. I also would like to thank all the Program Committee members for their careful and prompt review of submitted papers.

I would like to thank the Steering Committee Chair Professor Shi-Kuo Chang for his guidance and leadership throughout the organization of this conference. The assistance of the staff at KSI Research is also greatly appreciated, which made the review process smooth and timely.

Finally, I would like to thank you all for joining us in DMSVIVA2023, and really appreciate your participation and your desire to support the community year by year.

Walter Balzano, University of Naples, Italy; Program Chair

Table of Content

Session II

Notes: (S) denotes a short paper.

Keynote

Development of the Smart City Digital Twin of Florence

Dr. Marco Fanfani
University of Florence
Italy

Abstract

Digital Twins of Smart Cities are today a fundamental technology since they can offer a virtual context replicating a real city with high fidelity exploiting 3D building models enriched with contextual information coming from IoT sensors, heatmaps, analytic services, etc. Such solutions can provide a fundamental tool for decision-makers and stakeholders which can not only observe in real-time the status of the city but also perform analysis and simulations. In this talk we will discuss the steps needed to build a Smart City Digital Twin, from the definition of requirements to the 3D modeling and map construction, to its distribution through an interactive web interface, showing also how to embed information coming from sensors and analytics. We will focus on the case study of Florence implemented in the Snap4City platform.

About the Speaker

Dr. Marco Fanfani obtained his Ph.D. in Information Engineering in 2015 from the University of Florence, Italy. He is a senior research fellow at the DISIT Lab of the Department of Information Engineering of the University of Florence. His research interest includes IoT, digital twin, computer vision, 3D reconstruction, and virtual and augmented reality. He worked on several national and international R&D projects. He is cooperating on the development of the digital twin of the city of Florence, implemented in the Snap4City platform, presented at the 2022 Smart City Expo World Congress, and awarded as the best paper in the past DMSVIVA2022 conference.

Formal Specification and Model Checking of Raft Log Replication in Maude

Takanori Ishibashi
School of Information Science
Japan Advanced Institute of Science and Technology (JAIST)
Ishikawa, Japan
takanori.ishibashi@jaist.ac.jp

Kazuhiro Ogata
School of Information Science
Japan Advanced Institute of Science and Technology (JAIST)
Ishikawa, Japan
ogata@jaist.ac.jp

***Abstract*—Raft is a popular distributed consensus protocol and is used to build highly available and strongly consistent services in the industry. Using Maude, we formally specify the log replication in Raft and conduct model checking to check whether the protocol enjoys the Log Matching Property and the State Machine Safety Property. The Log Matching Property is that if two logs contain an entry with the same index and term, then the logs are identical in all entries up through the given index. The State Machine Safety Property is that if any two servers have applied two entries to their state machines at a same index, the two entries must be always the same. Our model checking experiments show that the protocol enjoys the properties under the condition that we limit the length of the server's log and the number of servers.**

***Index Terms*—state machines, invariant properties, rewriting logic, search command**

I. Introduction

Distributed consensus protocols, such as Raft [1], play a critical role in modern computing systems. These systems require high availability and strong consistency. The correctness of the systems is highly dependent on the correctness of Raft, and therefore, Raft provides significant functionality, and a bug in Raft can have tremendous effects; however, it is difficult to implement correctly distributed consensus protocols and to test their correctness. A proof assistant, such as Coq [2], allows us to prove that Raft enjoys the desired properties, but the time required for the investigation would probably be very long.

In this paper, we concentrate on the log replication, which is one of the basic mechanisms in Raft. We formally specify the log replication in Raft using Maude, which is a rewriting logic-based specification/programing language. We model check with Maude that Raft enjoys the Log Matching Property and the State Machine Safety Property, which are the properties that Raft is expected to guarantee. The Log Matching Property is that if two logs contain an entry with the same index and term, then the logs are identical in all entries up through the given index. The State Machine Safety Property is that if any two servers have applied two entries to their state machines at a same index, the two entries must be always the same. Our model checking experiment shows that Raft enjoys the two properties. The formal specification of the Raft log replication in Maude is available online[1].

The contribution of the work described in the present paper is to demonstrate how the log replication in Raft is formally specified in Maude and model checked with the Maude and to show that Raft enjoys the two properties under the condition that the length of the server's log and the number of servers are limited. We assume that a server in a Raft cluster conducts unexpected operations, which is different from the log replication in Raft. A server failure can result not only in a simple shutdown, but also in incorrect behavior. It is preferable to be able to handle the latter as well. Our model checking experiments also show that servers except for a server that conducts unexpected operations enjoy the two properties. To our knowledge, no study has examined this.

Diego Ongaro's original description of Raft [3] showed a formal specification of Raft in TLA+ [4] but did not show how to model check invariants of Raft with the TLA model checker. The description reported that using the TLA model checker on the specification was attempted but was abandoned because this approach did not scale well to larger models. The formal verification of the safety properties of the Raft was conducted using the Verdi framework for distributed systems verification [5]. This formal verification in Verdi proves some invariants, and consists of approximately 50,000 lines of Coq [2].

II. Preliminaries

Raft [1] is a distributed consensus protocol. Raft divides a distributed consensus problem into two independent sub-problems: leader election and log replication[2]. In leader election, Raft chooses at most one leader in each logical time called a term. There is one and only one leader in a Raft cluster in regular operations and all the other servers are then followers. In log replication, the leader accepts requests from clients, saves such requests in its log, and forwards them

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant JP24H03370.

DOI reference number: 10.18293/DMSVIVA23-010

[1]https://github.com/11Takanori/raft-maude

[2]Diego Ongaro et al. discussed that Raft divides a distributed consensus problem into three independent sub-problems: leader election, log replication, and safety [1]. In this paper, we consider that Raft divides the problem into two independent sub-problems: leader election and log replication, and safety is a requirement to be satisfied by the two subproblems.

to all the other servers. On receipt of such requests, each server saves them in its log. When the leader receives positive replies for a client request from the majority of servers, it commits (or consents to) the request. Each server has a state machine in which clients' requests are processed. When a follower receives a message saying that a client request has been committed, the follower commits the clients' request up to the client request (inclusive).

We describe the log replication in a bit more detail. The leader serves the client's request and appends it to its log. To replicate log entries, the leader sends appendEntries request messages to each of the other servers in the Raft cluster. Each server that has received the appendEntries request message responds to the message by sending a appendEntries response message to the leader. The appendEntries request message includes the following:

- term - the leader's term
- leaderId - the leader's ID
- prevLogIndex - the index of the log entry immediately preceding new ones
- prevLogTerm - the term of the log entry immediately preceding new ones
- entries - the log which contains leader's term and the client request messages
- leaderCommit - the leader's commit index

The appendEntries response message includes the following:

- term - the receiver's term
- success - a boolean value that means whether the receiver appends the new log entry to its log or not

If the follower that receives appendEntries request message finds an entry in its log with the same *prevLogIndex* and *prevLogTerm*, and the follower's term is not bigger than the leader's term, then the follower appends the new log entry to its log; otherwise, the follower refuses the new log entry. When the log entry has been replicated on a majority of the servers in the Raft cluster, the leader applies the log entry to its state machine (called commit). If the leader's commitIndex is greater than the commitIndex of the follower, the leader's commitIndex and the follower's index of the last new entry are compared and the follower commits at a smaller one. If the followers' log is inconsistent with the leader's log, the leader decrements nextIndex and retries the appendEntries request message. The leader maintains a nextIndex for each follower, which is the index of the next log entry the leader will send to that follower. In log replication, Raft is expected to guarantee the Log Matching Property and the State Machine Safety Property. These properties are discussed in the introduction. Raft is expected to guarantee that each of all properties is true under all non-Byzantine conditions, including network delays, duplication, partitions, message loss, and reordering.

A state transition system[3] is $\langle S, I, T \rangle$, where S is a set of states, $I \subseteq S$ is the set of initial states, $T \subseteq S \times S$ is a binary relation over states. Each element $(s, s') \subseteq T$ is called a state transition from *s* to *s'* and *T* is called the state transitions. There are multiple possible ways to express states. In this paper, we express a state as a braced associative-commutative collection of name-value pairs, where a name may have parameters. Associative-commutative collections are called soups according to the nomenclature of the Maude community, and name-value pairs are called observable components. That is, a state is expressed as a braced soup of observable components. We use the juxtaposition operator as the constructor of soups. Let oc_0, oc_1, oc_2 be observable components, and then $oc_0\, oc_1\, oc_2$ is the soup of those three observable components. A state is expressed as $\{oc_0\, oc_1\, oc_2\}$. *T* is specified in terms of rewrite rules. A rewrite rule starts with the keyword *rl*, followed by a label enclosed with square brackets and a colon, two pattern connected with $=>$, and ends with a period. A conditional rewrite rule starts with the keyword *crl* and have a condition following the keyword if before a period. The following is a form of a conditional rewrite rule: *crl [lb] : l => r if* ... $/\backslash\ c_i\ /\backslash$ ... where lb is a label given to the rule and c_i is part of the condition, which may be an equation $lc_i = rc_i$. The negation of $lc_i = rc_i$ could be written as *(lci =/= rci) = true*, where = *true* could be omitted. If the condition ... $/\backslash\ c_i\ /\backslash$ ... holds under some substitution σ, $\sigma(l)$ can be replaced with $\sigma(r)$.

[3]It may be called a state machine but because what are called state machines are used by Raft, we use the terminology "state transition system" in this paper.

III. Formal specification of the log replication

To formalize the log replication in Raft as a state transition system, we use the following observable components.

- *(term[s]: t)* - *s* is a server ID. *t* is a term. This means that the term of a server *s* is *t*. For each server s participating in a Raft cluster, an instance of this observable component is used.
- *(role[s]: r)* - *s* is a server ID. *r* is a role: leader or follower. This means that the role of a server *s* is *r*. For each server participating in a Raft cluster, an instance of this observable component is used.
- *(log[s]: l)* - *s* is a server ID. *l* is a list of log entries. This means that a server *s* has a log *l*. A server *s* appends log entries to the log *l*. For each server a participating in Raft cluster, an instance of this observable component is used.
- *(commitIndex[s]: ci)* - *s* is a server ID. *ci* is an index of highest log entry known to be committed. This means that index *ci* is the highest index at which a server s has committed log entries. For each server s participating in a Raft cluster, an instance of this observable component is used.
- *(nextIndex[s0][s1]: ni)* - *s0* and *s1* are server IDs. *ni* is an index of the next log entry that has been sent to s1 by s0. This means that a server *s0* will next send a server *s1* a log entry at nextIndex *ni*. A leader maintains a nextIndex for each follower, which is the index of the next log entry the leader will send to the follower. For each follower of a leader, an instance of this observable component is used.
- *(matchIndex[s0][s1]: mi)* - *s0* and *s1* are server IDs. *mi* is an index of the highest log entry such that s0 knows

that s1 has its replication of the log entry. This means that index *mi* is the highest index at which server *s0* has sent server *s1* a log entry. For each follower of a leader, an instance of this observable component is used.
- *(servers: ss)* - *ss* is a soup of server IDs. This maintains the IDs of all servers participating in a Raft cluster. One instance is only used and *ss* never changes.
- *(clientRequests: c)* - *c* is the messages that a client sends to a Raft cluster. One instance is only used. When a client sends a Raft cluster a message, the message is deleted from clientRequests.
- *(network: n)* - *n* is a soup of messages. This expresses the network with which the servers participating in a Raft cluster exchange messages. One instance is only used. A message that has been put into *n* is never deleted, which expresses that a message may be duplicated. Although a message is never deleted from the network, it would be possible for a server to never receive a message addressed to the server, which expresses that a message may be lost.

When there are three servers *s0*, *s1* and *s2* that participate in a Raft cluster, the initial state defined *init* is as follows:

```
{(term[s0]: 1) (term[s1]: 1) (term[s2]: 1)
 (role[s0]: leader) (role[s1]: follower)
 (role[s2]: follower)
 (log[s0]: empty) (log[s1]: empty) (log[s2]: empty)
 (commitIndex[s0]: 0) (commitIndex[s1]: 0)
 (commitIndex[s2]: 0)
 (nextIndex[s0][s1]: 1) (nextIndex[s0][s2]: 1)
 (matchIndex[s0][s1]: 0) (matchIndex[s0][s2]: 0)
 (servers: (s0 s1 s2)) (clientRequests: (cr0 cr1))
 (network: empty)} .
```

In the initial state, each value is as follows:
- the term of each sever is 1
- the role of the server *s0* is a leader
- the role of the server *s1* and the server *s2* is a follower
- the list of log entries that has been appended by each server is empty (meaning that each server has not yet appended any log entries to its log)
- the index that each server has committed is 0 (meaning that each server has not yet committed any log entries)
- the index of the next log entry that the server *s0* will send to the server *s1* and the server *s2* is 1
- the highest index at which the server *s0* has sent the server *s1* and the server *s2* a log entry is 0 (meaning that server *s0* has note yet known to be replicated on server *s1* and server *s2*)
- the soup of the servers that participate in a Raft cluster is *s0 s1 s2* because we suppose that the three servers participate in the Raft cluster
- the soup of the client request messages is *cr0 cr1* because we suppose that the client sends two messages to the Raft cluster
- the network is empty (meaning that no message has been put into the network)

The log replication in Raft is specified as three rewrite rules for each server: appendEntries, handleAppendEntriesRequest, and handleAppendEntriesResponse. The rewrite rules use the following Maude variables:
- *OCs* is a variable of observable components soups
- *S0*, *S1*, and *S2* are variables of server IDs
- *Ss* is a variable of server ID soups
- *T*, *U*, and *PLT* are variables of terms
- *R* and *R0* are variables of server roles
- *CI* and *LCI* are variables of commitIndex
- *MI*, *MI1*, and *MI2* are variables of matchIndex
- *NI*, *NI1*, and *NI2* are variables of nextIndex
- *PLI* is a variable of index of log entry immediately preceding new ones
- *PLT* is a variable of term of log entry immediately preceding new ones
- *Ls* and *L* are variables of log entries soups
- *CR* is a variable of client request messages
- *CRs* is a variable of client request message soups
- *NW* is a variable of message soups
- *AEReq* is a variable of AppendEntries request messages

The rewrite rule *appendEntries* is defined as follows:

```
rl [appendEntries] :
{(term[S0]: T) (role[S0]: leader)
 (commitIndex[S0]: CI) (log[S0]: Ls)
 (servers: Ss) (clientRequests: (CR CRs))
 (network: NW) OCs} =>
{(term[S0]: T) (role[S0]: leader)
 (commitIndex[S0]: CI)
 (log[S0]: Ls[length(Ls) + 1] := log(T, CR))
 (servers: Ss) (clientRequests: CRs)
 (network: (NW mkAppendEntriesRequests(
   S0, appendEntriesRequest(T, S0,
   length(Ls), term(Ls[length(Ls)]),
   log(T, CR), CI), Ss - S0))) OCs} .
```

The rewrite rule says that when a server *S0* is a leader and there exists a client request *CR*, *S0* puts an appendEntries request message in the network addressed to all the other servers. *appendEntriesRequest(T, S0, length(Ls), term(Ls[length(Ls)]), log(T, CR), CI)* is the body of the appendEntries request message. The first argument is *S0*'s term. The second argument is the leader's server ID. The third argument is the index of the log entry immediately preceding new ones (called a prevLogIndex). The fourth argument is the term of the log entry immediately preceding new ones (called a prevLogTerm). The fifth argument is the log which contains *S0*'s term and the client request messages. The sixth argument is *S0*'s commit index. $Ss - S0$ is the soup of server IDs obtained by deleting the server ID *S0* from the soup *Ss* of server IDs. *mkAppendEntriesRequests(S0, AEReq, Ss')* makes the appendEntries request message whose body is *AEReq* and that is addressed to all server IDs in *Ss'*.

The rewrite rule *handleAppendEntriesRequest* is defined as follows:

```
crl [handleAppendEntriesRequest] :
{(term[S0]: T) (role[S0]: R)
 (commitIndex[S0]: CI) (log[S0]: Ls)
 (network: (msg(S1, S0,
   appendEntriesRequest(U, S1, PLI, PLT,
    log(U, CR), LCI)) NW)) OCs} =>
{(term[S0]: if U > T then U else T fi)
 (role[S0]: if U >= T then follower else R fi)
 (commitIndex[S0]:
  if LCI > CI then min(LCI, length(L))
              else CI fi) (log[S0]: L)
 (network: (msg(S0, S1, appendEntriesResponse(
  (if U > T then U else T fi), B,
```

```
   appendEntriesRequest(U, S1, PLI, PLT,
    log(U, CR), LCI)))
  msg(S1, S0, appendEntriesRequest(
      U, S1, PLI, PLT, log(U, CR), LCI)) NW)) OCs}
if length(Ls) < 3
/\ B := U >= T and ((Ls[PLI] =/= null
        and term(Ls[PLI]) == PLT) or (PLI == 0))
/\ L := if B and Ls[PLI + 1] == null
        then Ls[PLI + 1] := log(U, CR)
        else (if (CI < PLI)
              and (length(Ls) =/= PLI
              or term(Ls[PLI]) =/= PLT)
              then Ls[: PLI] else Ls fi) fi .
```

When there exists *msg(S1, S0, appendEntriesRequest(U, S1, PLI, PLT, log(U, CR), LCI))* in the network and the length of the server *S0*'s log is less than 3, the following are conducted. If the server *S0*'s current term *T* is less than or equal to *U* ($U >= T$) and the server *S0* has an entry at prevLogIndex whose term matches prevLogTerm or the leader has not yet appended any log entries to its log ((Ls[PLI] =/= null and term(Ls[PLI]) == PLT) or (PLI == 0)), then *B* is true. If *B* is true and the server *S0* has not yet had a log entry at the index $PLI + 1$ (*B and Ls[PLI + 1] == null*), then the sever *S0* appends the log entry to the log. If the commitIndex of the server *S0* is less than prevLogIndex ($CI < PLI$) and an existing log entry of the server S0 conflicts with a new one (*length(Ls) =/= PLI or term(Ls[PLI]) =/= PLT*), then *S0* deletes the existing entry and all that follow it. If the commitIndex of the leader is greater than the commitIndex of the server *S0*, then the commitIndex of the leader and the index of last new entry are compared and the commitIndex of the server *S0* is updated by the smaller one. If $U > T$, then the server *S0*'s current term becomes *U* and *msg(S0, S1, appendEntriesResponse(U, B, appendEntriesRequest(U, S1, PLI, PLT, log(U, CR), LCI))* is put into the network; otherwise the server *S0*'s current term remains the same and *msg(S0, S1, appendEntriesResponse(T, B, appendEntriesRequest(U, S1, PLI, PLT, log(U, CR), LCI))* is put into the network. If $U >= T$, then the server S0 becomes a follower. Note that *msg(S1, S0, appendEntriesRequest(U, S1, PLI, PLT, log(U, CR), LCI)* is not deleted from the network. This is because a message may be duplicated.

The rewrite rule *handleAppendEntriesResponse* is defined as follows:

```
crl [handleAppendEntriesResponse] :
{(term[S0]: T) (role[S0]: R)
 (commitIndex[S0]: CI)
 (nextIndex[S0][S1]: NI1) (matchIndex[S0][S1]: MI1)
 (matchIndex[S0][S2]: MI2) (log[S0]: Ls)
 (network: (msg(S1, S0,
  appendEntriesResponse(U, B, AEReq)) NW)) OCs} =>
{(term[S0]: if U > T then U else T fi) (role[S0]: R0)
 (commitIndex[S0]:
  if replicatedCount(N, (MI MI2)) >= majority
  and T == term(Ls[N]) and R0 == leader
  and N > CI then N else CI fi)
 (nextIndex[S0][S1]: NI) (matchIndex[S0][S1]: MI)
 (matchIndex[S0][S2]: MI2) (log[S0]: Ls)
 (network: if (not B and U <= T and R0 == leader)
  then (msg(S0, S1, appendEntriesRequest(
          T, S0, PI, term(Ls[PI]), Ls[NI],
          leaderCommit(AEReq)))
       msg(S1, S0, appendEntriesResponse(
          U, B, AEReq)) NW)
  else (msg(S1, S0, appendEntriesResponse(
          U, B, AEReq)) NW) fi) OCs}
if MI := if B then max(prevLogIndex(AEReq) + 1, MI1)
         else MI1 fi
/\ NI := if B then MI + 1 else max(sd(NI1, 1), 1) fi
/\ PI := sd(NI, 1) /\ N := prevLogIndex(AEReq) + 1
/\ R0 := if U > T then follower else R fi .
```

When there exists *msg(S1, S0, appendEntriesResponse(U, B, AEReq))* in the network, the following are conducted. If *B* that is included in *appendEntriesResponse* is true, *S0*'s matchIndex for *S1* is updated with the greater of *prevLogIndex(AEReq) + 1* and *MI1*, where *prevLogIndex(AEReq) + 1* is the prevLogIndex (included in its appendEntriesRequest) plus 1 and *MI1* is server *S0*'s matchIndex for server *S1*, and server *S0*'s nextIndex for server S1 is updated with server *S0*'s matchIndex for server *S1* plus 1; otherwise, *S0*'s nextIndex for *S1* is updated with the greater of the *S0*'s nextIndex for *S1* minus 1 and 1 (the minimum value for nextIndex is 1). It is wrong to simply increment the matchIndex and the nextIndex because an *appendEntriesResponse* may be duplicated and the server *S0* may receive the appendEntriesResponse multiple times. Let *N* be the previous index plus 1. If the server *S0*'s matchIndex for server *S1* or the server *S0*'s matchIndex for server *S2* is greater than or equal to *N*, the term of server *S0*'s log at index *N* is server *S0*'s current term, the role of server *S0* is a leader and *N* is greater than server *S0*'s commitIndex, then server *S0*'s commitIndex is updated by *N*. If *B* is false, *U* is less than or equal to *T* and the role of server *S0* is a leader, then *msg(S0, S1, appendEntriesRequest(T, S0, PI, term(Ls[PI]), Ls[NI], leaderCommit(AEReq)))* is put into the network, meaning that server *S0* sends appendEntriesRequest message with the log entry at the previous nextIndex because of log inconsistency between server *S0* and server *S1*. If $U > T$, then the *S0*'s term becomes *U*, *S0* becomes a follower. Note that *msg(S1, S0, appendEntriesResponse(U, B, AEReq))* is not deleted from the network due to as mentioned beforehand.

IV. Model checking the log replication

The experimental environment used is SUSE Linux Enterprise Server 15 SP1 installed on a computer with a 2.8GHz 16 core processor and 1.5 TB memory. The computer is maintained by Research Center for Advanced Computing Infrastructure, JAIST. We are allowed to keep on using the computer up to one week in a row for each job, a model checking experiment for us. If a model checking experiment is not completed in one week, the job is killed. Under condition that the length of server's log is less than 3 and the number of servers is 3, we conducted some model checking experiments.

A. Model checking under assumptions that no server conducts unexpected operations

We first confirmed that logs are replicated correctly in our specification by using the following Maude command:

```
search [1] in RAFT : init =>*
{(role[S0:ServerID]: leader)
 (log[S0:ServerID]: L0:Logs)
 (log[S1:ServerID]: L1:Logs)
 (log[S2:ServerID]: L2:Logs)
 (commitIndex[S0]: 2)
 (commitIndex[S1]: 1) (commitIndex[S2]: 1)
 (matchIndex[S0][S1]: 2) (matchIndex[S0][S2]: 2)
```

```
 (nextIndex[S0][S1]: 3) (nextIndex[S0][S2]: 3) OCs}
such that length(L0:Logs) == 2
  and length(L1:Logs) == 2 and length(L2:Logs) == 2 .
```

Maude finds a solution for the search command. The solution says that there exists a path from the initial state leading to a state in which all servers have two log entries, the leader commits at index 2, the other servers commit at index 1, the leader's matchIndex for the other servers is 2, and the leader's nextIndex for the other servers is 3. This means that two log entries are replicated correctly. Note that the result dose not say that log entries will be eventually replaced correctly. There is a path along which two log entries have not been replicated.

We use Maude to check that Raft satisfies the Log Matching Property or not by using the following search command:

```
search [1] in RAFT : init =>*
{(log[S0:ServerID]: L0:Logs)
 (log[S1:ServerID]: L1:Logs)
 (matchIndex[S0][S1]: I:Nat) OCs}
such that L0:Logs[I:Nat] =/= null
 and L1:Logs[I:Nat] =/= null
 and (term(L0:Logs[I:Nat])
  == term(L1:Logs[I:Nat]))
 and ((term(L0:Logs[sd(I:Nat, 1)])
  =/= term(L1:Logs[sd(I:Nat, 1)]))
 or (value(L0:Logs[sd(I:Nat, 1)])
  =/= value(L1:Logs[sd(I:Nat, 1)]))) .
```

The result returned by Maude for the search command is as follows:

```
No solution. states: 2805  rewrites: 1049267 in 488ms
cpu (486ms real) (2150137 rewrites/second)
```

That is to say, Maude did not find any state in which two logs contain a log entry with the same index and term, and the logs contain a different log entry in all entries up through the index. Consequently, we can conclude that Raft enjoys the Log Matching Property under the condition that the length of server's log is less than 3 and the number of servers is 3.

We use Maude to check that Raft satisfies the State Machine Safety Property or not by using the following search command:

```
search [1] in RAFT : init =>*
{(log[S0:ServerID]: L0:Logs)
 (log[S1:ServerID]: L1:Logs)
 (commitIndex[S0]: I:Nat)
 (commitIndex[S1]: I:Nat) OCs}
such that ((term(L0:Logs[I:Nat])
  =/= term(L1:Logs[I:Nat]))
 or (value(L0:Logs[I:Nat])
  =/= value(L1:Logs[I:Nat]))).
```

The result returned by Maude for the search command is as follows:

```
No solution. states: 2805  rewrites: 993397 in 448ms
cpu (451ms real) (2217404 rewrites/second)
```

That is to say, Maude did not find any state in which a server has applied a log entry at a given index to its state machine, and other servers apply a different log entry for the same index. Consequently, we can conclude that Raft enjoys the State Machine Safety Property under the condition that the length of server's log is less than 3 and the number of servers is 3.

B. Model checking under assumptions that a server conducts unexpected operations

We assume that a follower conducts unexpected operations in a Raft cluster. We define *SERVER-ID* module in the previous subsection as follows:

```
fmod SERVER-ID is
 sort ServerID .
 ops s0 s1 s2 : -> ServerID [ctor] .
endfm
```

In this subsection, we modify *SERVER-ID* module as follows:

```
fmod SERVER-ID is
 sorts ServerID BadServerID .
 subsort BadServerID < ServerID .
 ops s0 s1 : -> ServerID [ctor] .
 op s2 : -> BadServerID [ctor] .
endfm
```

This means *ServerID* and *BadServerID* are sorts, sort *BadServerID* is a subsort of sort *ServerID*, *s0*, *s1* and *s2* are constants, *s0*'s sort and *s1*'s sort are ServerID, and *s2*'s sort is *BadServerID*.

We add the following rewrite rule.

```
crl [badHandleAppendEntriesRequest] :
{(term[S2:BadServerID]: T) (role[S2:BadServerID]: R)
 (commitIndex[S2:BadServerID]: CI)
 (log[S2:BadServerID]: Ls)
 (network: (msg(S1, S2:BadServerID,
   appendEntriesRequest(U, S1, PLI, PLT,
   log(U, CR), LCI)) NW)) OCs} =>
{(term[S2:BadServerID]: if U > T then U else T fi)
 (role[S2:BadServerID]:
   if U >= T then follower else R fi)
 (commitIndex[S2:BadServerID]: length(L))
 (log[S2:BadServerID]: L)
 (network: (msg(S2:BadServerID, S1,
   appendEntriesResponse(
    (if U > T then U else T fi), true,
    appendEntriesRequest(U, S1, PLI, PLT,
     log(U, CR), LCI)))
  msg(S1, S2:BadServerID, appendEntriesRequest(
    U, S1, PLI, PLT, log(U, CR), LCI)) NW)) OCs}
if length(Ls) < 3 /\ L := Ls[PLI + 1] := log(U, crb) .
```

When the sort of sever ID is *BadServerID*, there exists *msg(S1, S2:BadServerID, appendEntriesRequest(U, S1, PLI, PLT, log(U, CR), LCI))* in the network and the length of the server *S0*'s log is less than 3, the following are conducted. The sever *S2* appends the log entry including client request *crb* and term *U* to the log. *crb* is the unexpected value and it is not the value that client sent. The commitIndex of server *S2* is updated by the length of server *S2*'s log. If $U > T$, then the server *S0*'s current term becomes *U* and *msg(S2:BadServerID, S1, appendEntriesResponse(U, true, appendEntriesRequest(U, S1, PLI, PLT, log(U, CR), LCI))* is put into the network; otherwise the server *S0*'s current term remains the same and *msg(S2:BadServerID, S1, appendEntriesResponse(T, true, appendEntriesRequest(U, S1, PLI, PLT, log(U, CR), LCI))* is put into the network. If $U >= T$, then the server *S0* becomes a follower. Note that *msg(S1, S2:BadServerID, appendEntriesRequest(U, S1, PLI, PLT, log(U, CR), LCI)* is not deleted from the network due to as already mentioned.

This rewrite rule shows an unexpected operation. How to replicate a log entry and how to commit are different from the log replication in Raft. The server *s2* whose sort is

BadServerID conducts the rewrite rule *handleAppendEntries-Request* or the rewrite rule *badHandleAppendEntriesRequest*. This is because sort *BadServerID* is a subsort of sort *ServerID*. That is to say, the server *s2* conducts both normal operations and unexpected operations.

We confirmed again that logs are replicated correctly in our modified specification by using the following search command:

```
search [1] in RAFT : init =>*
{(role[S0:ServerID]: leader)
 (log[S0:ServerID]: L0:Logs)
 (log[S1:ServerID]: L1:Logs)
 (log[S2:ServerID]: L2:Logs)
 (commitIndex[S0]: 2)
 (commitIndex[S1]: 1) (commitIndex[S2]: 1)
 (matchIndex[S0][S1]: 2) (matchIndex[S0][S2]: 2)
 (nextIndex[S0][S1]: 3) (nextIndex[S0][S2]: 3) OCs}
such that length(L0:Logs) == 2
  and length(L1:Logs) == 2 and length(L2:Logs) == 2 .
```

Maude finds a solution for the search command. The solution says that there exists a path from the initial state leading to a state in which all servers have two log entries, the leader commits at index 2, the other servers commits at index 1, the leader's matchIndex for the other servers is 2, and the leader's nextIndex for the other servers is 3. This means that two log entries are replicated correctly. Note that the result dose not say that log entries will be eventually replaced. There is a path along which two log entries have not been replicated.

We use Maude to check that Raft satisfies the Log Matching Property for the server *s0* and the server *s1* or not by using the following Maude command:

```
search [1] in RAFT : init =>*
{(log[s0]: L0:Logs) (log[s1]: L1:Logs)
 (matchIndex[s0][s1]: I:Nat) OCs}
such that L0:Logs[I:Nat] =/= null
 and L1:Logs[I:Nat] =/= null
 and (term(L0:Logs[I:Nat])
  == term(L1:Logs[I:Nat]))
 and ((term(L0:Logs[sd(I:Nat, 1)])
  =/= term(L1:Logs[sd(I:Nat, 1)]))
 or (value(L0:Logs[sd(I:Nat, 1)])
  =/= value(L1:Logs[sd(I:Nat, 1)]))) .
```

The result returned by Maude for the search command is as follows:

```
No solution. states: 24987
rewrites: 8900588 in 6364ms cpu
(6389ms real) (1398583 rewrites/second)
```

That is to say, Maude did not find any state in which the server *s0* and the server *s1* have a log entry with the same index and term, and the server *s0*'s log and the server *s1*'s log contain a different log entry in all entries up through the index. Consequently, we can conclude that Raft enjoys the Log Matching Property in servers which do not conduct unexpected operations under the condition that the length of server's log is less than 3 and the number of servers is 3.

We use Maude to check that Raft satisfies the State Machine Safety Property for the server *s0* and the server *s1* or not by using the following search command:

```
search [1] in RAFT : init =>*
{(log[s0]: L0:Logs) (log[s1]: L1:Logs)
 (commitIndex[s0]: I:Nat)
 (commitIndex[s1]: I:Nat) OCs}
such that ((term(L0:Logs[I:Nat])
  =/= term(L1:Logs[I:Nat]))
 or (value(L0:Logs[I:Nat])
  =/= value(L1:Logs[I:Nat]))).
```

The result returned by Maude for the search command is as follows:

```
No solution. states: 24987
rewrites: 8380677 in 5412ms cpu
(5417ms real) (1548536 rewrites/second)
```

That is to say, Maude did not find any state in which the server *s0* or the server *s1* has applied a log entry at a given index to its state machine, and the opposite server apply a different log entry for the same index. Consequently, we can conclude that Raft enjoys the State Machine Safety Property in servers which do not conduct unexpected operations under the condition that the length of server's log is less than 3 and the number of servers is 3.

Under the condition that the length of the server's logs is less than 4 and the number of servers is 3 and the condition that the length of the server's logs is less than 3 and the number of servers is 4, we conducted model checking experiments to check whether Raft enjoys the Log Matching Property or not, and the State Machine Safety Property or not. However, the model checking experiments took over a week to complete and the job running model checking was killed.

V. Conclusion

We reported that the log replication in Raft is formally specified in Maude and model checking experiments are conducted based on the formal specification. Our model checking experiments have said that logs are replicated correctly and Raft enjoys the Log Matching Property that if two logs contain an entry with the same index and term, then the logs are identical in all entries up through the given index and the State Machine Safety Property that if any two servers have applied two entries to their state machines at a same index, the two entries must be always the same. Under assumptions that a server conducts unexpected operations, which are different from the log replication in Raft, our model checking experiments have said that Raft enjoys the Log Matching Property and the State Machine Safety Property in servers that do not conduct unexpected operations.

References

[1] Diego Ongaro and John Ousterhout. 2014. In search of an understandable consensus algorithm. In Proceedings of the 2014 USENIX conference on USENIX Annual Technical Conference (USENIX ATC'14). USENIX Association, USA, 305-320. https://dl.acm.org/doi/10.5555/2643634.2643666

[2] Yves Bertot and Pierre Castéran. 2013. Interactive theorem proving and program development: Coq'Art: the calculus of inductive constructions. Springer Science & Business Media.

[3] Diego Ongaro. 2014. Consensus: Bridging Theory and Practice. Ph.D. Dissertation. Stanford University.

[4] Lamport, L. Specifying Systems, The TLA+ Language and Tools for Hardware and Software Engineers. Addison-Wesley, 2002.

[5] Doug Woos, James R. Wilcox, Steve Anton, Zachary Tatlock, Michael D. Ernst, and Thomas Anderson. 2016. Planning for change in a formal verification of the raft consensus protocol. In Proceedings of the 5th ACM SIGPLAN Conference on Certified Programs and Proofs (CPP 2016). Association for Computing Machinery, New York, NY, USA, 154-165. https://doi.org/10.1145/2854065.2854081

Automatic Identification and Extraction of Assumptions on GitHub

Chen Yang[†‡], Zinan Ma[†], Peng Liang[§¶*], Xiaohua Liu[†]
[†]School of Artificial Intelligence, Shenzhen Polytechnic, Shenzhen, China
[‡]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[§]School of Computer Science, Wuhan University, Wuhan, China
[¶]Hubei Luojia Laboratory, Wuhan, China
{yangchen, 21680260, lxh}szpt.edu.cn, liangp@whu.edu.cn

***Abstract*—In software development, due to the lack of knowledge or information, time pressure, complex context, and many other factors, various uncertainties emerge during the development process, leading to assumptions scattered in projects. Being unaware of certain assumptions can result in critical problems (e.g., system vulnerability and failures). The prerequisite of analyzing and understanding assumptions in software development is to identify and extract those assumptions with acceptable effort. In this paper, we proposed a tool (i.e., Assumption Miner) to automatically identify and extract assumptions on GitHub projects. To evaluate the applicability of Assumption Miner, we first presented an example of using the tool to mine assumptions from one large and popular deep learning framework project: the TensorFlow project on GitHub. We then conducted an evaluation of the tool. The results show that Assumption Miner can effectively identify and extract assumptions from the repositories on GitHub.**

***Index Terms*—Assumption, GitHub, Mining Software Repositories**

I. Introduction

Assumptions in the field of software development is a broad topic: different types of assumptions (e.g., requirement assumptions [1], design assumptions [2], and construction assumptions [3]) have been extensively discussed. For instance, in the early phases of software development, there could be many uncertain things. However, in order to meet the project business goals (e.g., schedule and deadlines), stakeholders have to work in the presence of such uncertainties; these uncertainties can lead to assumptions. In this paper, we advocate treating uncertainty and assumption as two different but related concepts: one way to deal with uncertainties is to make implicit or explicit assumptions (i.e., a thing that is uncertain, but accepted as true) [4].

The importance of assumptions and their management in software development has been highlighted in many studies and industrial cases. For example, Corbató [5] mentioned in his ACM Turing Award lecture that "*design bugs are often subtle and occur by evolution with early assumptions being forgotten as new features or uses are added to systems.*" Garlan et al. pointed out that incompatible assumptions in software architecture can cause architectural mismatch [6]. Lewis et al. also mentioned similar results in machine learning systems: since there are different types of stakeholders (e.g., data scientist, software engineer, and system user) of a machine learning system, they could make different but incompatible or invalid assumptions, leading to system misunderstanding, mismatch, etc [7]. In October 2018, Lion Air Flight 610 crashed 13 minutes after takeoff and killed all 189 people on board; In March 2019, Ethiopian Airlines Flight 302 crashed and ended another 157 lives. According to the reports from the government, one critical reason of the 737 MAX crashes is regarding not-well managed assumptions [8] [9]. In the report, they mentioned that the aircraft company made invalid assumptions about the critical system components. Specifically, the invalid assumptions regarding MCAS (Maneuvering Characteristics Augmentation System) are the root cause of the crashes. The report also pointed out the need of identifying and re-evaluating important assumptions in the system.

As evidenced from researchers and practitioners, stakeholders constantly make assumptions in their work [4] [10]. The assumptions can be further classified as two types: self-claimed assumptions (SCAs) and potential assumptions (PAs). Considering a sentence in a commit, pull request (PR), or issue of a GitHub repository, an SCA is that the sentence includes an assumption and the assumption is explicitly claimed using at least one of the assumption-related terms (i.e., "assumption", "assumptions", "assume", "assumes", "assumed", "assuming", "assumable", and "assumably"). For example, in the sentence: "*[tf/xla] fixup numbering of xla parameters used for aliasing previously, the xla argument parameter was incorrectly assumed to be corresponding to the index in the vector of 'xlacompiler::argument'*", it includes an SCA of assuming the XLA ARGUMENT parameter is corresponding to the index in the vector of XLACOMPILER::ARGUMENT and the developer claimed that it was an invalid assumption. A PA is that the sentence may include an assumption, i.e., it is a potential assumption that needs further confirmation from human experts. This definition covers various aspects, such as expectations, future events, possibilities, guesses, opinions, feelings, and suspicions, which can indicate PAs. We provide three examples to further explain the PA concept. For example, in the sentence: "*i think the right way to create demo tensorboard instances is to simply run a tensorboard in the cloud, rather than keep maintaining this mocked-out*

DOI reference number: 10.18293/DMSVIVA2023-071

backend.", it includes a PA regarding thoughts of the right way to create demo tensorboard instances. In another example: "*The system will not crash under heavy load*", the sentence describes a future state of the system, which is uncertain, and includes a PA. The third example is: "*both false and true outputs should be considered independently*". This sentence does not include assumption-related terms, but the sentence describes an expectation (i.e., "something should be"), which is a PA. After further confirmation by human experts, this PA can be transformed to an SCA or other types of software artifacts. Besides SCAs and PAs (which belong to explicit assumptions), there are also many implicit assumptions in projects (e.g., in stakeholders' heads or requiring reasoning). Identifying implicit assumptions at the sentence level is much more tricky than identifying SCAs and PAs, since there are no explicit clues in the sentences and stakeholders need to infer the sentences based on the context, which involves assumption reasoning. Identifying implicit assumptions is out of the scope of this study, and we treat it as future work.

Assumptions are related to many types of software artifacts, such as decisions, technical debt, and source code [11]. For example, in TensorFlow, there is an SCA: "*TODO: Looks like there is an assumption that weight has only one user. We should add a check here*", which induces a technical debt. Existing research on assumptions and their management in software development usually use experiments, surveys, and case studies to manually identify and extract assumptions through observation, questionnaires, interviews, focus groups, and documentation analysis [4]. Since such approaches have high costs (e.g., time and resources), the number of identified and extracted assumptions in those studies is often limited, leading to various problems of developing new theories, approaches, and methods of assumptions and their management in software development.

In this work, to overcome the issues of manually identifying and extracting assumptions in software development, we proposed a tool: Assumption Miner, which can be used to automatically identify and extract assumptions (i.e., SCAs and PAs) on GitHub repositories. With over 100 million developers, 4 million organizations, and 330 million repositories[1], GitHub is one of the most important sources for open source software development. Besides assumptions, the tool can also be easily extended to other research fields (e.g., identifying and extracting technical debt [12]).

How to access Assumption Miner. Assumption Miner is available at [2]. Users can register or use a *guest* account to login the tool. We also provided a deployment package for users who want to try the tool on their local environment [13]. Users can read and follow the instructions in the description for the deployment of the tool.

The remainder of the paper is organized as follows. Section II provides related work, Section III describes the details of Assumption Miner, Section IV presents an example of using the tool, Section V describes an evaluation of Assumption Miner, and Section VI concludes the paper with future directions.

[1]https://github.com/about/, accessed on 2023-04-21

[2]http://39.108.224.140

II. Related Work

In the field of assumptions and their management in software development, most assumptions are manually identified and extracted by researchers and practitioners. Landuyt and Joosen focused on assumptions made during the application of a threat modeling framework (i.e., LINDDUN), which allows the identification of privacy-related design flaws in the architecting phase [14]. They conducted a descriptive study with 122 master students, and the students identified and extracted 845 assumptions from the models created by the students. Yang et al. conducted an exploratory study of assumptions made in the development of nine popular deep learning frameworks (e.g., TensorFlow, Keras, and PyTorch) on GitHub [11]. They identified and extracted 3,084 assumptions from the code comments in over 50,000 files of the deep learning frameworks. Xiong et al. studied assumptions in the Hibernate developer mailing list, including their expression, classification, trend over time, and related software artifacts [15]. In their study, they identified and extracted 832 assumptions. Li et al. developed a machine learning approach [16] to identify and classify assumptions based on the dataset constructed by Xiong et al. [15], which can read the data (i.e., sentences) from the dataset (i.e., a .CSV file), preprocess the data (e.g., using NLTK and Word2Vec), train classifiers (e.g., Perception, Logistic Regression, and Support Vector Machines), and evaluate the trained classifiers (e.g., precision, recall, and F1-score). However, their approach is not specifically developed for PAs and SCAs and cannot mine assumptions from other sources (e.g., GitHub repositories).

Compared to the related work above, the tool (i.e., Assumption Miner) proposed in this work focuses on GitHub repositories, and can automatically collect data (e.g., issues, PRs, and commits) and identify and extract SCAs and PAs. Assumption Miner can also be easily extended to work with other repositories (e.g., Stack Overflow) or other types of software artifacts (e.g., technical debt).

III. Assumption Miner

Assumption Miner is composed of four modules: Repository Management, Data Collection, Data Extraction, and System Management, as shown in Fig. 1.

Repository Management includes (1) getting information of the repositories and their releases from the GitHub server, (2) searching and showing details of the repositories, (3) downloading source code of each release of the repositories, and (4) deleting all the data of specific repositories. Adding a repository using Assumption Miner requires users to enter the owner (e.g., tensorflow) and name (e.g., tensorflow) of the repository. When adding a repository, Assumption Miner first checks whether the repository exists in the MySQL database and the GitHub server. If all the checks pass, Assumption Miner gets information (e.g., URL, releases, and tags) of the

repository from the GitHub server and insert the data into the MySQL database.

Data Collection aims to (1) show the data models of Repository, Release, Tag, PR, Commit, and Issue, (2) search and show data collection information of repositories and collect issues, PRs, and commits based on the data models, (3) monitor data collection processes, and (4) show data collection history. The data models are predefined and currently cannot be changed by users. The basic information of each repository (e.g., its releases and tags) and the information of the data collection processes are stored in a MySQL database, while the data of issues, PRs, and commits are stored in a MongoDB database. When collecting data from the GitHub server using Assumption Miner, users can also set a time (default: 10 seconds) for automatically refreshing the data collection status (i.e., collecting, finished, and error). Assumption Miner uses cursors (i.e., issue cursor, PR cursor, and commit cursor) to record each batch of the data downloaded from the GitHub server, and therefore Assumption Miner supports continuing data collection after it has been stopped due to errors and exceptions (e.g., over the limits by GitHub).

Data Identification and Extraction is composed of three submodules: Assumption Extraction, Data Search, and Knowledge Graph. **Assumption Extraction** contains four functions: SCA Identification, SCA Extraction, PA Identification, and PA Extraction. In the ***SCA Identification*** function, we used a keyword-based search approach for identifying SCAs (i.e., word level), based on the *assumption* related search terms (i.e., *assumption*, *assumptions*, *assume*, *assumes*, *assumed*, *assuming*, *assumable*, and *assumably*) and the following search scope: (1) title, body, body of comments of issues, (2) title, body, body of comments of PRs, and (3) message of commits. Since the results from using the SCA Identification function are at the word level (i.e., highlighting the matched terms), in the ***SCA Extraction*** function, Assumption Miner provides support for locating, matching, and extracting the related sentences that include the matched terms (i.e., at the sentence level). As most of the description (e.g., description of an issue) is created and edited by GitHub users, there could be punctuation problems existing in the description (e.g., a sentence does not have a "." or "." is used in the source code, such as "a.b"), which may lead to errors in the separation of the sentences. Therefore, we manually identified the patterns of SCA description from the projects on GitHub, and implemented the patterns in the SCA Extraction function to extract SCA sentences. The outputs of SCA Extraction contain a set of data items as shown in Table I and Table II. Moreover, there are sentences without using *assumption* related keywords, but could act as assumptions (we call them potential assumptions, PAs). Though PAs are not SCAs, they can be further reviewed by stakeholders and transformed to SCAs (as the inputs for SCA extraction). Therefore, Assumption Miner also provides support for identifying and extracting such assumptions through the ***PA Identification*** and ***PA Extraction*** function, which complements SCA identification and extraction. The first author manually collected and labeled 35,855 sentences from the issues, PRs, and commits of multiple repositories (e.g., Keras and Theano), following the guidelines of assumption identification proposed in [4]. Then the other authors reviewed the results and reached a consensus with the first author. After assumption collection and labeling, we constructed a dataset for PAs, fine-tuned a deep learning model based on ALBERT (a lite BERT, which architecture is based on BERT), and trained and adjusted a classification model for PA identification [17]. The reason of choosing ALBERT is because ALBERT is one of the most powerful language models, which can achieve good performance with fewer parameters (compared to BERT) in many tasks, such as the binary single-sentence classification task [17]. Since we are using deep learning models and identifying PAs at the sentence level, the identification process could be rather slow on CPUs (e.g., it may take hours/days on our server to identify PAs, depending on the amount of data to be processed). Therefore, we used a queue (i.e., a waiting list with a first-in first-out strategy) on the server to manage the tasks of identifying PAs. In the PA Extraction function, Assumption Miner organizes the PAs (sentences) identified from the PA Identification function into a file (similar to SCA Extraction), and users can download the file for further review.

TABLE I: Data Items of Issues and PRs in SCA Extraction

Data Item	Description
owner	The owner of the repository
repo_name	The name of the repository
author	The author of the issue or PR
title	The title of the issue or PR
state	The state of the issue or PR
url	The URL of the issue or PR
SCA	The SCA sentence in the issue or PR

TABLE II: Data Items of Commits in SCA Extraction

Data Item	Description
owner	The owner of the repository
repo_name	The name of the repository
author_name	The author of the issue or PR
message	The message of the issue or PR
url	The URL of the issue or PR
SCA	The SCA sentence in the issue or PR

Data Search aims to search and show specific issues, PRs, and commits based on keywords. In data search, Assumption Miner requires users to specify which repository (e.g., *TensorFlow*), data type (i.e., *issue, PR*, and *commit*), search scope (e.g., *title*), and keywords (e.g., *assume*) to search. For example, for issues of the TensorFlow project, a search scope can be *title body comments.body*, which means that Assumption Miner will search data within the scope of the title of the issues, the body of the issues, and the body of the comments of the issues in the TensorFlow project. For keywords, Assumption Miner supports AND (i.e., using double quotation marks, e.g., *"assume" "software"*) and OR (i.e., without quotation marks, e.g., *assume software*). The search terms are highlighted in the search results. If the

description of a data item is too long, users can click on the "detail" button to see the full information of the item.

Knowledge Graph supports both traditional knowledge graph and dynamic knowledge graph. Assumption Miner provides three dimensions (i.e., release, month, and day) to construct the timeline of the data. For each repository, Assumption Miner creates and connects entities according to the timeline and their states. For example, a PR can be published, merged, and closed, and therefore Assumption Miner creates three connected entities if they are within a timeline.

System Management supports user registration, login, and logoff, and provides access control and system logs.

The architecture of Assumption Miner is shown in Fig. 2. When a user clicks on a menu or button, the Web component organizes the data, generates a request, and sends it to the Controller component through the Python Interface component. The Controller component analyzes the request: (1) If the request is regarding getting data from the GitHub server, the Controller component organizes the data and calls the functions in the GitHub Service component. The GitHub Service component reads the GitHub configuration (stored in the system) and organizes queries based on the predefined data models. Then the GitHub Service component sends requests to the GitHub server, gets responses from the GitHub server, analyzes the responses, and sends back the data to the Controller component. (2) If the request is regarding interacting with the MySQL or the MongoDB database, the Controller component organizes the data and calls the functions in the Data Service component. The Data Service component further organizes the data and calls the functions in the DAO (Data Access Object) component, which implements the interaction with the MySQL or the MongoDB database. The DAO component reads the database configuration (stored in the system), communicates with the databases, gets the results from the databases, and sends them back to the Data Service component. The Data Service component sends the data getting from the databases back to the Controller component. (3) If the request is regarding using the trained model (based on ALBERT) to identify PAs, the Controller component organizes the data and calls the functions in the Data Service component. The Data Service component preprocesses the data, loads the trained model, and uses the model to identify PAs. The results are then sent back to the Controller component. Finally, the Controller component organizes the data getting from GitHub, the MySQL database, the MongoDB database or the deep learning model, sends the data back to the Web component through the Python Interface component, and then the Web component shows the results to the user.

IV. Using Assumption Miner

In this section, we walk the usage of Assumption Miner through an example: the TensorFlow project on GitHub [3]. TensorFlow is one of the most popular deep learning frameworks, which is widely used in many deep learning systems and application domains. The TensorFlow project on GitHub started in 2015, having 188 releases [4], 146,893 commits[5], 22,887 PRs [6], and 37,074 issues [7] till April 2023. Users need to register an account and set a personal access token of GitHub [8] when using Assumption Miner. The token is used to access the GitHub Application Programming Interface (API), since Assumption Miner needs to communicate with the GitHub API to get data (e.g., issues, PRs, and commits). We also provide a default token for Assumption Miner users. However, since GitHub has limitations in place to protect against excessive or abusive calls to GitHub servers (e.g., the rate limit is 5,000 points per hour and individual calls cannot request more than 500,000 total nodes)[9], using the default token may lead to errors in data collection because of these limitations. After registration of the Assumption Miner account, users can login Assumption Miner with the account. Below is the process of using Assumption Miner to identify and extract assumptions from the TensorFlow project on GitHub.

Create the TensorFlow repository. Users need to click on the Repository Management module, then click on the "Add" button, enter the owner as "tensorflow" and the name as "tensorflow", and click on the "Save" button to create the TensorFlow repository on Assumption Miner. For each release of a repository, Assumption Miner provides users a link to download the source code in the Repository Management module (this is an optional step). Then users can use tools such as Visual Studio Code and PyCharm to further browse the code and search assumptions in the code.

Collect issues, PRs, and commits on TensorFlow. After the TensorFlow repository is created on Assumption Miner, users can further use the Data Collection module to collect issues, PRs, and commits of the TensorFlow repository. Users can start multiple tasks simultaneously, but this could cause errors because of the limitation by GitHub.

Identify and Extract SCAs on TensorFlow. In the Assumption Extraction submodule of the Data Extraction module, users need to select the *TensorFlow* repository and a data type, and click on the "SCA Identification" button. Assumption Miner will show the results and highlight all the SCAs. Users can further extract the data to a CSV file to construct a dataset by clicking on the "SCA Extraction" button. The first line in the CSV file is the title, indicating the repository and type (i.e., issue, PR, and commit) of the extracted data. If users want to search data on the TensorFlow repository (optional step), they can use the Data Search submodule of the Data Extraction module, by selecting the repository (e.g., "TensorFlow") and the data type (e.g., "issue") and specifying

[3]https://github.com/tensorflow/tensorflow

[4]https://github.com/tensorflow/tensorflow/releases (accessed on 2023-04-21)

[5]https://github.com/tensorflow/tensorflow (accessed on 2023-04-21)

[6]https://github.com/tensorflow/tensorflow/pulls?q=is%3Apr (accessed on 2023-04-21)

[7]https://github.com/tensorflow/tensorflow/issues?q=is%3Aissue (accessed on 2023-04-21)

[8]http://www.m58.link/cxaNX

[9]https://docs.github.com/en/graphql/overview/resource-limitations

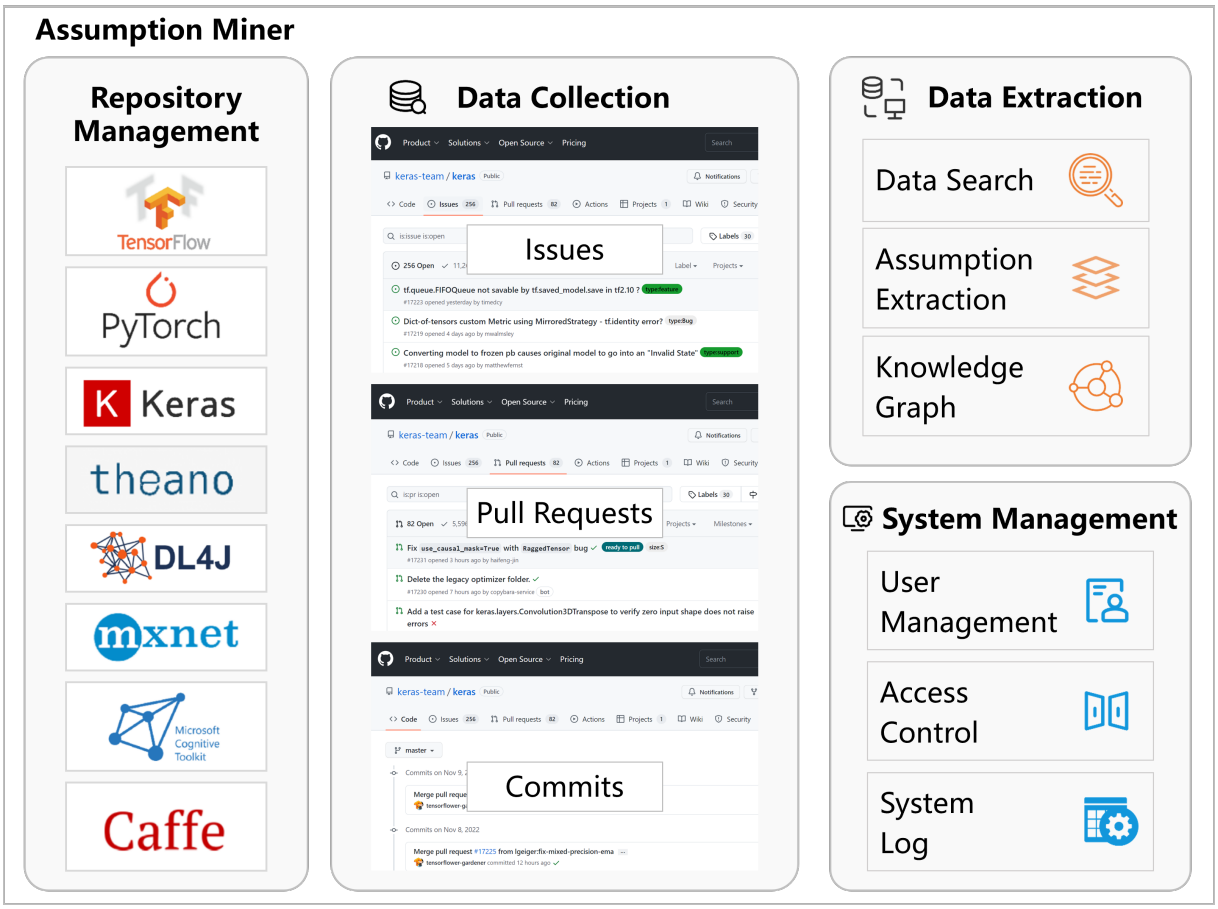


Fig. 1: Modules of Assumption Miner

the search scope (e.g., "title") and the search term (e.g., "assumption").

Identify and Extract PAs on TensorFlow. In the Assumption Extraction submodule of the Data Extraction module, users need to select the *TensorFlow* repository and a data type, and click on the "PA Identification" button. Assumption Miner will show the results and highlight all the sentences that may include a PA. After PA identification, users can click on the "PA Extraction" button to download the results of PA identification for further review.

Generate a knowledge graph of the assumptions on TensorFlow. This step is optional. Users need to select the *TensorFlow* repository and a dimension (i.e., *release*, *month*, or *day*) to construct a knowledge graph of assumptions based on the chosen dimension.

The aforementioned results can be further used in various context. For example, through identifying assumptions, users may better understand what was assuming in a certain project and deal with such uncertainty in their future work.

V. Evaluation of Assumption Miner

We conducted an evaluation on data collection, SCA identification, SCA extraction, PA identification, and PA extraction, as shown in Fig. 3. The output of data collection is the input of SCA identification and PA identification, the output of SCA identification is the input of SCA extraction, and the output of PA identification is the input of PA extraction.

A. Evaluation of Data Collection

We conducted an evaluation of using Assumption Miner to collect issues, PRs, and commits on seven GitHub repositories: Caffe, CNTK, Theano, DeepLearning4J (DL4J), MXNet, Keras, and TensorFlow, regarding the completeness and performance of Assumption Miner on data collection. We first created an issue data model (including *repository name*, *title*, *ID*, *author*, *URL*, *labels*, *state*, *body*, and *comments*), a PR data model (including *repository name*, *owner*, *title*, *ID*, *author*, *URL*, *labels*, *state*, *body*, *comments*, and *reviews*), and a commit data model (including *repository name*, *owner*, *OID*, *author name*, *author email*, *committed date*, *URL*, and *message*). The data items of each data model were selected from the GraphQL API on GitHub[10].

The configuration of the server we used for the evaluation of the data collection is: (1) CPU: Intel(R) Xeon(R) Platinum 8255C CPU @ 2.50GHz, 2 cores, (2) Memory: 2GB, (3) Hard Disk: 40GB SSD, (4) Operation System: Linux VM-20-4-centos 3.10.0-1160.45.1.el7.x86_64 #1 SMP. The results

[10]https://docs.github.com/en/graphql/reference/objects

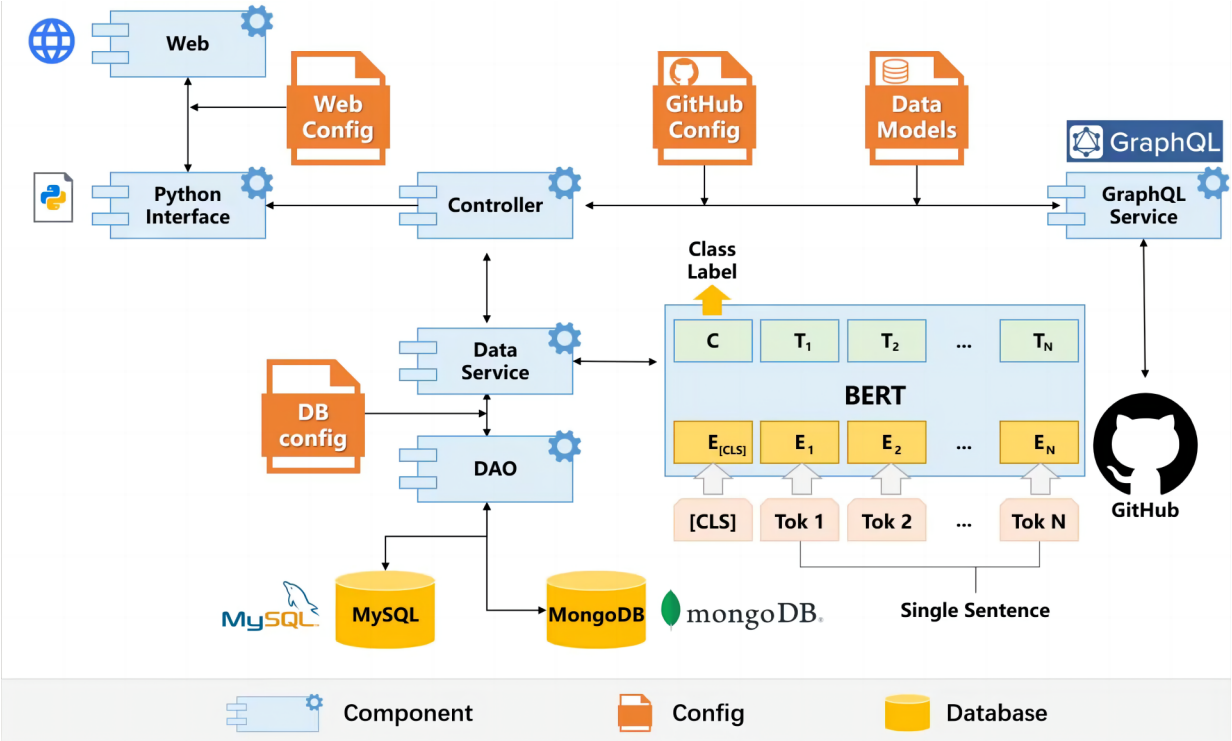


Fig. 2: Architecture of Assumption Miner

of using Assumption Miner to collect data are shown in Table III.

The results show that Assumption Miner can effectively collect issues, PRs, and commits of the repositories on GitHub. Certain repositories (e.g., TensorFlow) may be frequently updated, and there is a need to construct a mechanism for continuously collecting data.

B. Evaluation of SCA Identification

After data collection (as mentioned in Section V-A), we used the collected data (i.e., issues, PRs, and commits) of the Keras and TensorFlow repository to conduct an evaluation on identifying SCAs using Assumption Miner (i.e., the SCA Identification function). The first author manually checked the identified results to classify them as SCAs or non-SCAs. The evaluation results of SCA identification are shown in Table IV.

The count of the identified SCAs could be larger than the search results (e.g., count of messages in the commits of the Keras repository), since each issue, PR, or commit may include multiple SCAs. For example, an issue [11] of Keras mentions: "*Assume we are trying to learn a sequence to sequence map. For this we can use Recurrent and TimeDistributedDense layers. Now assume that the sequences have different lengths. We should pad both input and desired sequences with zeros, right? But how will the objective function handle the padded values? There is no choice to pass a mask to the objective function. Won't this bias the cost function?*", which includes two SCAs: "assume we are trying to learn a sequence to sequence map" and "assume that the sequences have different lengths".

Since Assumption Miner used a keyword-based (i.e., the *assumption* related terms) search approach for SCA identification, it could go wrong in certain context (e.g., a variable in a code snippet named *assume*). Moreover, we also found that certain SCAs lack details. For example, an issue [12] of Keras mentioned: "*strict enforcement of user input assumptions, and raising of helpful error messages.*" However, we cannot understand what exactly the user input assumptions are. These SCAs need to be further processed by Assumption Miner (e.g., add warnings in the results).

Regarding the performance of the SCA Identification process, we used an Aliyun server [13] with the following configuration: (1) Server type: ecs.s6-c1m2.xlarge, (2) CPU: 4 cores (vCPU), (3) Memory: 8GB, (4) Hard Disk: 40GB SSD, (5) Operation System: Windows Server 2022, and conducted an evaluation on the issues, PRs, and commits of Keras and Tensorflow. The results are shown in Table V.

The results show that Assumption Miner can correctly identify 94.92% SCAs (1,961 SCAs out of 2,066 SCAs) from the issues, PRs, and commits of the Keras and TensorFlow repository. Certain variables and functions named as for example *assume* exist in issues, PRs, and commits, which needs to be further processed by Assumption Miner.

[11] https://github.com/keras-team/keras/issues/395

[12] https://github.com/keras-team/keras/issues/1174

[13] https://www.alibabacloud.com/en

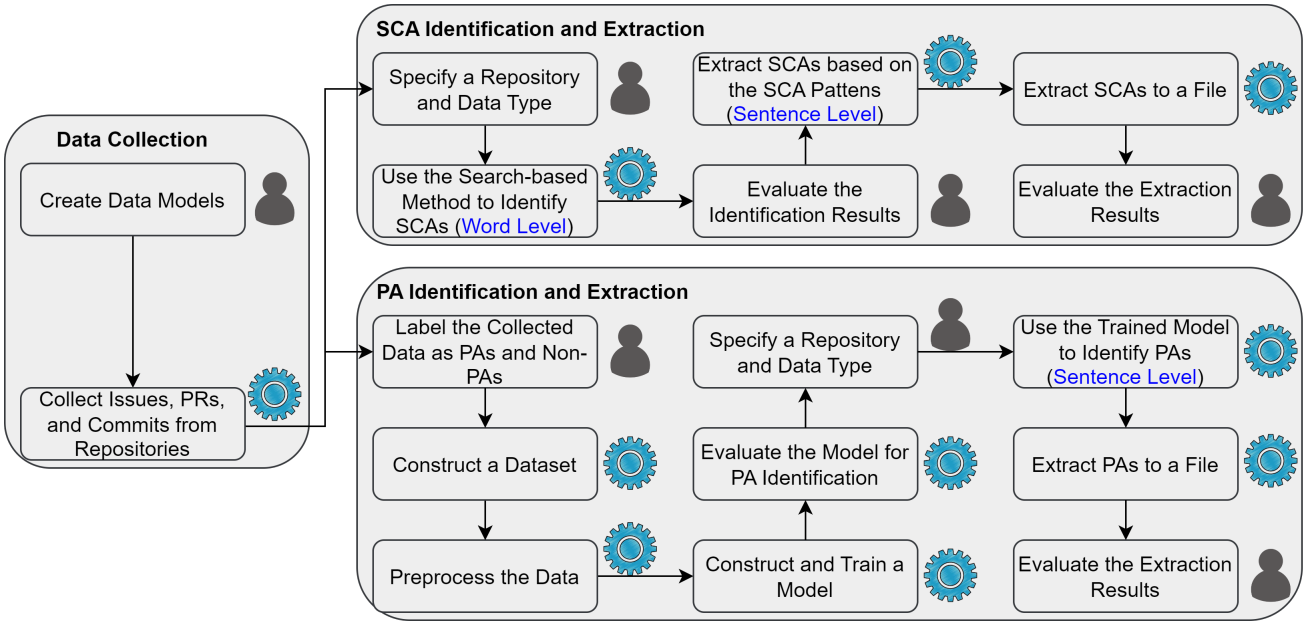


Fig. 3: Evaluation of Assumption Miner

TABLE III: Completeness and Performance of Data Collection of using Assumption Miner

Repository	Data Items on GitHub	Collected Data Items	Data Type	Time (s)	Access Date
Caffe	4,786	4,786	Issue	400	2022/11/4
Caffe	2,238	2,238	PR	152	2022/11/4
Caffe	4,156	4,156	Commit	51	2022/11/4
CNTK	3,288	3,288	Issue	228	2022/11/4
CNTK	557	557	PR	27	2022/11/4
CNTK	16,117	16,117	Commit	181	2022/11/4
Theano	2,671	2,671	Issue	246	2022/11/4
Theano	4,114	4,114	PR	198	2022/11/4
Theano	28,127	28,127	Commit	335	2022/11/4
DL4J	5,652	5,652	Issue	290	2022/11/7
DL4J	4,185	4,185	PR	140	2022/11/7
DL4J	2,606	2,606	Commit	25	2022/11/7
MXNet	9,532	9,532	Issue	465	2022/11/8
MXNet	11,090	11,090	PR	551	2022/11/8
MXNet	11,893	11,893	Commit	145	2022/11/7
Keras	11,518	11,518	Issue	519	2022/11/8
Keras	5,670	5,670	PR	208	2022/11/8
Keras	7,493	7,493	Commit	65	2022/11/8
TensorFlow	35,966	35,966	Issue	2,402	2022/11/9
TensorFlow	22,119	22,119	PR	1,105	2022/11/9
TensorFlow	138,366	138,366	Commit	1,295	2022/11/9

C. Evaluation of SCA Extraction

We further evaluated whether Assumption Miner (i.e., the SCA Extraction function) can correctly extract SCAs (i.e., at the sentence level) using the identification results (i.e., 1961 identified SCAs) of the Keras and TensorFlow repository from V-B. The first author manually checked the extracted results to classify them as correct extraction and missed extraction. The results are shown in Table VI. As an example, there are 298 SCAs in the body of Keras issues. Assumption Miner correctly extracted 290 of the SCAs, but missed 8 SCAs.

The results show that Assumption Miner can correctly extract 97.55% SCAs (1,913 SCAs out of 1,961 SCAs) from the issues, PRs, and commits of the Keras and TensorFlow repository. Certain structures of the issues, PRs, and commits (e.g., "assume" and "assumption" exist in one sentence) may lead to errors in SCA extraction, which needs further investigation and improvements.

D. Evaluation of PA Identification

For PAs, we manually labeled 35,855 sentences from the issues, PRs, and commits of multiple repositories (e.g., Keras and Theano), and constructed a dataset, containing a training set and a test set with a data proportion of 8:2 from the labeled sentences. We created a vocabulary and tokenized the data from the dataset based on the vocabulary. Then we constructed the deep learning model (based on ALBERT), trained the

TABLE IV: Results of Identifying SCAs using Assumption Miner

Repository	Data Type	Search Field	Search Results	Identified SCAs	Misidentification
Keras	Issue	title	3	3	0
Keras	Issue	body	253	298	13
Keras	PR	title	3	3	0
Keras	PR	body	59	65	0
Keras	Commit	message	10	11	0
TensorFlow	Issue	title	13	13	0
TensorFlow	Issue	body	662	797	47
TensorFlow	PR	title	5	5	0
TensorFlow	PR	body	136	150	6
TensorFlow	Commit	message	567	616	39
		Total	1,711	1,961	105

TABLE V: Performance of the SCA Identification Process

Repository	Data Type	Identification Field	Time (s)
Keras	Issue	title, body, comments.body	7.287
Keras	PR	title, body, comments.body	0.650
Keras	Commit	message	0.378
TensorFlow	Issue	title, body, comments.body	7.419
TensorFlow	PR	title, body, comments.body	0.632
TensorFlow	Commit	message	0.681

TABLE VI: Results of Extracting SCAs using Assumption Miner

Repository	Data Type	Extraction Field	SCAs	Correct	Missed
Keras	Issue	title	3	3	0
Keras	Issue	body	298	290	8
Keras	PR	title	3	3	0
Keras	PR	body	65	65	0
Keras	Commit	message	11	11	0
TensorFlow	Issue	title	13	13	0
TensorFlow	Issue	body	801	772	29
TensorFlow	PR	title	5	5	0
TensorFlow	PR	body	150	146	4
TensorFlow	Commit	message	616	609	7
		Total	1,961	1,913	48

model for 50,000 epochs with a batch size of 32 and a learning rate of 2e-5.

Regarding the performance of the PA identification process, we used the same configuration used in the evaluation of SCA identification (see Section V-B), and conducted an evaluation on the issues, PRs, and commits of Keras and Tensorflow. The results are shown in Table VII.

TABLE VII: Performance of the PA Identification Process

Repository	Data Type	Identification Field	Time
Keras	Issue	title, body, comments.body	11h 12m 58s
Keras	PR	title, body, comments.body	2h 28m 6s
Keras	Commit	message	24m 4s
TensorFlow	Issue	title, body, comments.body	47h 2m 34s
TensorFlow	PR	title, body, comments.body	12h 16m 59s
TensorFlow	Commit	message	10h 19m 44s

Using GPU would significantly improve the performance of the PA identification process. As an example, we used NVIDIA GeForce RTX 3060 Ti to run the PA identification process on Keras. The results are 16 minutes 52 seconds (compared to 11 hours 12 minutes 58 seconds using CPU) on issues, 3 minutes 51 seconds (compared to 2 hours 28 minutes 6 seconds using CPU) on PRs, and 43 seconds (compared to 24 minutes 4 seconds using CPU) on commits of Keras.

The results show that the best accuracy of identifying PAs using Assumption Miner on the test set is 0.9451 on Epoch 22,000. Considering repositories may have various context (e.g., different development policies), there is a need to further extend the dataset to include more data from the repositories, which will help to improve the generalization ability of Assumption Miner in identifying PAs.

VI. Conclusions

Assumptions and their management are important in software development. The prerequisite of analyzing and understanding assumptions in software development is to identify and extract those assumptions with acceptable effort. To this end, we proposed Assumption Miner to automatically identify and extract assumptions on GitHub projects. Besides providing a running example of using Assumption Miner on the TensorFlow project, we also evaluated the performance of Assumption Miner, and the results show that Assumption Miner can effectively identify and extract assumptions from the repositories on GitHub. Assumption Miner can be potentially used for the research topics regarding assumptions and their management in software development, such as assumption making, evolution, evaluation, and reasoning.

For future work, the following aspects of Assumption Miner can be further optimized: (1) There is a need to construct a mechanism for continuously collecting data using Assumption Miner. (2) The identification of SCAs and PAs can be further optimized (e.g., develop new deep learning models, construct a larger dataset and train deep learning models based on the dataset). (3) Certain patterns of the issues, PRs, and commits (e.g., a variable named “assume”) may lead to incorrect SCA identification and extraction, which can be further addressed in Assumption Miner. (4) Assumptions are related to various types of software artifacts (e.g., requirements, design decisions, and technical debt). Automatically recovering the relationships between assumptions and such artifacts in Assumption Miner is a promising future direction. (5) Besides SCAs and PAs, there are many implicit assumptions in projects, which should also be identified and extracted in the future.

ACKNOWLEDGMENTS

This work is funded by Shenzhen Polytechnic with Grant No. 6022312043K, State Key Laboratory for Novel Software Technology at Nanjing University with Grant No. KFKT2022B37, the National Natural Science Foundation of China (NSFC) with Grant No. 62172311, and the Special Fund of Hubei Luojia Laboratory.

REFERENCES

[1] C. B. Haley, R. C. Laney, J. D. Moffett, and B. Nuseibeh, "Using trust assumptions with security requirements," *Requirements Engineering*, vol. 11, no. 2, pp. 138—151, 2006.

[2] R. Roeller, P. Lago, and H. van Vliet, "Recovering architectural assumptions," *Journal of Systems and Software*, vol. 79, no. 4, pp. 552–573, 2006.

[3] M. M. Lehman and J. F. Ramil, "Rules and tools for software evolution planning and management," *Annals of SE*, vol. 11, no. 1, p. 15–44, 2001.

[4] C. Yang, P. Liang, and P. Avgeriou, "Assumptions and their management in software development: A systematic mapping study," *Information and Software Technology*, vol. 94, no. 2, pp. 82–110, 2018.

[5] F. J. Corbató, "On building systems that will fail," *Communications of the ACM*, vol. 34, no. 9, p. 72–81, 1991.

[6] D. Garlan, R. Allen, and J. Ockerbloom, "Architectural mismatch: Why reuse is still so hard," *IEEE Software*, vol. 26, no. 4, p. 66–69, 2009.

[7] G. A. Lewis, S. Bellomo, and I. Ozkaya, "Characterizing and detecting mismatch in machine-learning-enabled systems," in *Proceedings of the 1st Workshop on AI Engineering: Software Engineering for AI (WAIN)*, Madrid, Spain, 2021, pp. 133–140.

[8] "Status of the Boeing 737 MAX - hearing before the subcommittee on aviation of the committee on transportation and infrastructure house of representatives," U.S. Government Publishing Office, 2019. [Online]. Available: https://www.govinfo.gov/content/pkg/CHRG-116hhrg37277/pdf/CHRG-116hhrg37277.pdf

[9] "Final committee report - the design, development, and certification of the Boeing 737 MAX," The Committee on Transportation and Infrastructure, 2020. [Online]. Available: https://www.edpierson.com/final-737-max-report-for-public-release

[10] C. Yang, P. Liang, and P. Avgeriou, "A survey on software architectural assumptions," *Journal of Systems and Software*, vol. 113, no. 3, pp. 362–380, 2016.

[11] C. Yang, P. Liang, L. Fu, and Z. Li, "Self-claimed assumptions in deep learning frameworks: An exploratory study," in *Proceedings of the 25th International Conference on Evaluation and Assessment in Software Engineering (EASE)*, Trondheim, Norway, 2021, pp. 139–148.

[12] Z. Li, P. Avgeriou, and P. Liang, "A systematic mapping study on technical debt and its management," *Journal of Systems and Software*, vol. 101, pp. 193–220, 2015.

[13] C. Yang, Z. Ma, P. Liang, and X. Liu, "Deployment package of Assumption Miner," 2023. [Online]. Available: https://doi.org/10.5281/zenodo.7725115

[14] D. V. Landuyt and W. Joosen, "A descriptive study of assumptions made in linddun privacy threat elicitation," in *Proceedings of the 35th Annual ACM Symposium on Applied Computing (SAC)*, Brno, Czech Republic, 2020, pp. 1280–1287.

[15] Z. Xiong, P. Liang, C. Yang, and T. Liu, "Assumptions in OSS development: An exploratory study through the hibernate developer mailing list," in *Proceedings fo the 25th Asia-Pacific Software Engineering Conference (APSEC)*, Nara, Japan, 2018, pp. 455–464.

[16] R. Li, P. Liang, C. Yang, G. Digkas, A. Chatzigeorgiou, and Z. Xiong, "Automatic identification of assumptions from the hibernate developer mailing list," in *Proceedings fo the 26th Asia-Pacific Software Engineering Conference (APSEC)*, Putrajaya, Malaysia, 2019, pp. 394–401.

[17] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bertt for self-supervised learning of language representations," in *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, Virtual Conference, 2020.

An interval RSP-based ensemble model for big data analysis

Wenzhu Cai, GengYuan Ao, YiGang Lin, Mark Junjie Li
College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
caiwenzhu2021@email.szu.edu.cn, jj.li@szu.edu.cn

***Abstract*—Ensemble learning for big data has been successful in machine learning and has great advantages over other learning methods. The ensemble model based on Random Sample Partition (RSP) is a prominent method of it. Although the RSP data blocks have the consistent probability distribution function as the whole data, there is some uncertainty in prediction results due to the non-overlapping data between blocks. In this paper, we propose a novel interval ensemble model based on RSP named Inr-RSP, which maps prediction results to interval-valued data by interval modeling and then uses the IAA aggregation method to convert the interval-valued data into fuzzy sets to get a more accurate and stable final result. The experimental classification results from four real datasets also show that the performance of this model is better than that of the traditional RSP ensemble model. And the IAA method usage has a stronger ability to capture uncertainty in prediction than the common majority voting method.**

***Index Terms*—Big data analysis, Ensemble learning, Random Sample Partition, Interval Agreement Approach**

I. Introduction

In recent years, due to the popularity of emerging technologies such as the Internet of Things (IoT), social media, and mobile devices, the scale of data has exploded. Faced with such massive amounts of data, how to efficiently process and analyze them has become an urgent problem to be solved. One of the biggest challenges in big data analysis is how to perform complex computing tasks within a given amount of computing resources. Previously, divide and conquer was the main strategy for big data analysis and calculation, which divide data into small subsets and then processed the subsets independently [1]. The MapReduce [2] and Spark [3], two distributed programming models, are also based on this strategy to process massive data. However, due to the iterative operation, the execution efficiency decreases, and the above models are limited by available memory resources in the calculation and analysis [4]. Therefore, the memory issue becomes a problem for big data analysis.

This problem is mainly alleviated by sampling techniques. Traditional sampling methods such as simple random sampling [5], stratified sampling [6], reservoir sampling [7], and the Record-Level Sampling (RLS) of the Hadoop Distributed File System (HDFS) [8] in distributed architectures, are all based on records. It becomes time-consuming and limited by memory in big data because selecting records with equal probability requires scanning the entire data. Ensemble learning is a common approach when using sampling techniques by dividing the data into many subsets or fitting multiple models using different algorithms, which typically improves the predictive performance of data mining and machine learning algorithms [9]. Nevertheless, the traditional ensemble models including Bagging [10] and Boosting [11] methods can not avoid the bottlenecking of memory resources when using the whole large dataset. Salman et al. propose an appropriate analysis model for large-scale datasets call Random Sampling Partition (RSP), which stores data as ready-to-use blocks of non-overlapping random samples [12]. The generation of RSP blocks is an offline operation, and each block has the consistent probability distribution with the whole data, thus providing the possibility of using a few blocks to approximate the whole big data without the limit of memory.

The existing RSP model generates the RSP blocks using the two-stage data processing (TSDP) [13] algorithm and then obtains the approximate result by processing each block respectively. Although RSP blocks have the consistent probability distribution with the big data, there is some uncertainty in the prediction results due to the non-overlapping of the data between the blocks. The common aggregation strategy is majority voting [14], but it does not consider the effect of the interval values. Besides, the number of learning models is determined by the number of learning models, which is not flexible.

In this paper, we propose an interval ensemble learning model based on RSP named Inr-RSP, which takes into account the uncertainty of the prediction results using interval modeling and uses the Interval Agreement Approach (IAA) to aggregate the final result. Experimental results show that the Inr-RSP model can achieve more accurate and robust classification with minimal information loss. Meanwhile, it presents that a few RSP blocks are enough to achieve the performance of the entire blocks and the number of learning models can be independent of the number of blocks which reduces model costs.

II. Related Works

Big data Sampling is a technology that extracts a sample set from a big dataset to facilitate data processing and analysis. The distribution of the sample data is important to the machine learning models. In the case of random sampling, the distribution of the predicted sampling is similar to that of the overall data. Common sampling methods include Bernoulli Sampling

DOI reference number: 10.18293/DMSVIVA2023-089

[15], Simple Random Sampling [5], Stratified Sampling [6], and so on [16]. Bernoulli Sampling [15] is to randomly select a single sample with moderate probability from the total with variable sample size and prone to sample bias. Simple Random Sampling [5] takes a lot of work when the data size is large or the distribution is more dispersed. Stratified Sampling [6] provides greater statistical precision and reduces sampling error. Similarly, the Bootstrap [17] method requires a large number of replicate samples and traversing the full data each time, which requires large enough memory resources. The RSP model divides large data into ready-to-use disjoint blocks whose distribution is consistent with that of the entire dataset. The use of RSP models can build ensemble models with fewer data, solving the problems of high computation and memory limitation [12].

The aggregation functions of ensemble learning are methods of combining multiple predictions into a final prediction result. Some of the most classic methods are majority voting [14], weighted voting [18], and stacking [19]. Majority voting [14] is the most common and effective method. Papers [20]–[22] applied the majority voting method to ensemble learning in different applications, and the results of the studies indicated that the majority voting method had a nice performance. In recent years, fuzzy theory has been used to deal with uncertain data, and interval-valued aggregation functions based on fuzzy theory have been proposed for ensemble learning [23]. Paper [24]–[27] proposed interval-valued aggregation functions to capture the uncertainty of data and applied them to ensemble learning. In particular, the Interval Agreement Approach (IAA) [24] converts interval-valued data into fuzzy sets. The IAA method addresses the limitations of the Interval Approach (IA) [28] and the Enhanced Interval Approach (EIA) [29] which only consider fuzzy sets of limited types and cannot handle uncertain intervals. It considers the minimal assumptions of interval data and does not rely on data preprocessing and outlier removal.

III. Preliminaries

In this section, we succinctly review the Random Sample Partition data model and briefly describe the Interval Agreement Approach.

A. Random Sample Partition (RSP)

Random Sample Partition is a distributed data model to facilitate block-level sampling and support big data analysis [12]. In this model, the statistical properties of the data set are preserved in a group of small disjoint data blocks as ready-to-use random samples (RSP blocks) from the entire data. Each RSP block has consistent probability distribution with the whole big data, allowing local results on different data blocks to approximate the global results on the whole big data. Also, it can address the limitation of memory and high computing cost in large-scale data.

With the RSP model, a partitioning of $\mathbb{D}$ into k non-overlapping random sample data blocks $T = \{D_1, D_2, ..., D_k\}$ in advance is represented as RSP blocks if:

- $\bigcup_{i=1}^{k} D_i = \mathbb{D}$
- $D_i \cap D_j = \emptyset,\ \ where\ i, j \in \{1, 2, ..., k\}\ \ and\ i \neq j$
- $E\left[F_i\left(x\right)\right] = \mathbb{F}\left(x\right),\ \ where\ i \in \{1, 2, ..., k\}$

where $F_i(x)$ is the sample probability distribution function of a random variable x in D_i. Accordingly, each block of T is called an RSP block of $\mathbb{D}$. Selecting an RSP block from T equals directly extracting random samples from $\mathbb{D}$. To analyze large-scale data, using such Block-Level Sampling is more efficient than Record-Level Sampling because it not requires scanning the entire data.

The RSP-based ensemble model for big data analysis uses a few selected RSP blocks to obtain approximate results. First, a block-level sample is selected from the RSP. Second, a sequential algorithm is applied parallel to each selected RSP block. Third, the outputs of these blocks are combined to produce an approximate result for the entire data (i.e., the majority voting in a classification task or the average response in a regression task). The ensemble process for the classification task is shown in Figure 1.

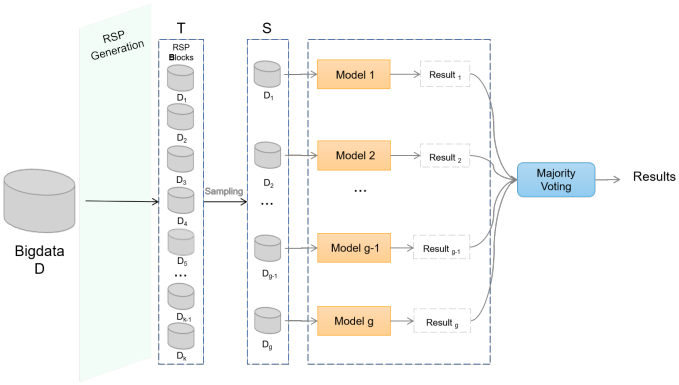


Fig. 1. The RSP-based ensemble model for big data analysis

B. Interval Agreement Approach (IAA)

The Interval Agreement Approach is a novel approach to generating fuzzy sets from interval-valued data and is accurately modeled by aggregating collective information captured by intervals [24]. An interval is denoted as $\bar{A} = [l_{\bar{A}}, r_{\bar{A}}]$, where $l_{\bar{A}}$ shows the left endpoint and $r_{\bar{A}}$ represents the right endpoint. Let $\mathcal{A} = \left\{\bar{A}_1, ..., \bar{A}_n\right\}$ be a set of intervals and a Type-1 Fuzzy Set (T1 FS) named A in IAA. The membership function μ_A of A is defined as:

$$\begin{aligned}\mu_A = {} & y_1 / \bigcup_{i_1=1}^{n} \bar{A}_{i_1} \\ & + y_2 / \left(\bigcup_{i_1=1}^{n-1} \bigcup_{i_2=i_1+1}^{n-1} \left(\bar{A}_{i_1} \cap \bar{A}_{i_2}\right) \right) \\ & + \dots \\ & + y_n / \left(\bigcup_{i_1=1}^{1} \dots \bigcup_{i_n=n}^{n} \left(\bar{A}_{i_1} \cap \dots \cap \bar{A}_{i_n}\right) \right)\end{aligned} \tag{1}$$

where $y_i = i/n$. Equation(1) represents the common notation of membership for fuzzy sets and / refers to the degree of membership rather than division. That means a value of μ_A shows the number of that value within all the intervals in $\mathcal{A}$. When the y_i is equal to 1, it indicates that all intervals are intersected.

There are two ways to simplify equation(1). One is that the membership of any value x can be calculated as the count of intervals which x contained like

$$\mu_A(x) = \frac{1}{n}\sum_{i=1}^{n}\mu_{\bar{A}_i}(x)$$
$$\textbf{where } \mu_{\bar{A}_i}(x) = \begin{cases} 1 & l_{\bar{A}_i} \le x \le r_{\bar{A}_i} \\ 0 & \text{else} \end{cases} \tag{2}$$

The other way to show the membership function is to subtract the number of left endpoints less than x in $\mathcal{A}$ from the number of right endpoints in $\mathcal{A}$ less than x as

$$\mu_A(x) = \frac{1}{n}\left(\sum_{i=1}^{n}\left(l_{\bar{A}_i} \le x\right) - \sum_{i=1}^{N}\left(r_{\bar{A}_i} \le x\right)\right) \tag{3}$$

Thus, A Type-1 Fuzzy Set can be generalized over $\mu_A(x)$.

IV. Proposed Model

In this section, we introduce the new interval RSP-based ensemble model named Inr-RSP which uses interval modeling and the IAA aggregation method to capture the uncertainty of prediction results and decrease the information loss. The main process of Inr-RSP is shown in Figure 2.

A. Generate RSP

Let $\mathbb{D}$ be a multivariate data set of N records and M features where N is large. A partitioning of $\mathbb{D}$ into k small disjoint data blocks $\{D_1, D_2, ..., D_k\}$ is regarded as a generation of RSP. The two-stage data processing (TSDP) algorithm for generation is as [13]:

- Sequentially cut $\mathbb{D}$ into p non-overlapping subsets called a partition of $\mathbb{D}$. Each subset has the same size with n records. Randomize each subset into i.i.d and cut it into an RSP of k parts independently to generate P data blocks.
- From each RSP block, select its corresponding RSP block, for 1 to k, to generate a new data block. Repeat this merging operation k times to generate a new partition $\{D_1, D_2, ..., D_k\}$, which is an RSP of $\mathbb{D}$.

The RSP model generates ready-to-use non-overlapping data blocks with consistent probability distribution of the entire data. It only needs to be executed once, which achieves Write-Once-Use-Many-Times(WOUM) strategy.

B. RSP Blocks Sampling

In this part, select g blocks from RSP data blocks $T = \{D_1, D_2, ..., D_k\}$ without replacement to form a sample set S as

$$S = \{D_1, D_2, ..., D_g\}$$

where $g \le k$. Thus, memory and communication costs depend on g, not k. The sampled RSP blocks are the same as the samples of the whole big data used for the following big data analysis.

C. Build Different Models

According to the analysis task, the base model can choose different learning models for the same task or one learning model with different parameters. For example, different base models, e.g. decision tree, support vector machine, and logistic regression, can be used if the task is classification.

D. Generate Uncertain Intervals and Aggregate

The key idea of the proposed model is capturing uncertainty by uncertain intervals from different data samples and models. Firstly, the selected RSP blocks are processed in different built models. In this part, intermediate results $\{Result_{i\_j}\}_{j=1}^{g}$ can be generated for each model i. Then, to avoid the influence of outliers on the results, we used Tukey's Test to process the intermediate result. Consider DL_i as $Q1_i - k * 1.5$ and UL_i as $Q3_i + k * 1.5$, the uncertain interval for each model i is shown as:

$$I_i = [DL_i, UL_i] \tag{4}$$

where the $Q1_i$ is the first quartile of $\{Result_{i\_j}\}_{j=1}^{g}$ and $Q3_i$ is the third quartile of $\{Result_{i\_j}\}_{j=1}^{g}$ for model i. Also, k is the difference between $Q3_i$ and $Q1_i$.

As mentioned before, IAA is used to generate a Type-1 Fuzzy Set (T1 FS), which is able to capture variation in the opinion of a particular decision model and divergence between the individual views of a group of decision models. Using uncertain intervals I_i and equation(2), a T1 FS is defined as:

$$A = \{((l_i, r_i), u_i)\}_{i=1}^{z} \tag{5}$$

where l_i is the left point, r_i is the right point and u is the membership function value of regions.

E. Defuzzification of Fuzzy Sets

There are many defuzzification methods to calculate the centroid of the Type-1 Fuzzy Sets. In this part, the computation approach of [30] is used to acquire the centroid as follows:

$$c = \frac{u_1 * (l_1 + r_1) + u_2 * (l_2 + r_2) + \cdots + u_z * (l_z + r_z)}{2(u_1 + u_2 + \cdots + u_z)} \tag{6}$$

where c is the centroid which will be applied as the final result. For binary classification, if the centroid is equal to or upper than 0.5, the final class will be class zero, and if it's not, it is class one. Similarly, the multiclass classification result is the primary class, and the other case is the secondary class.

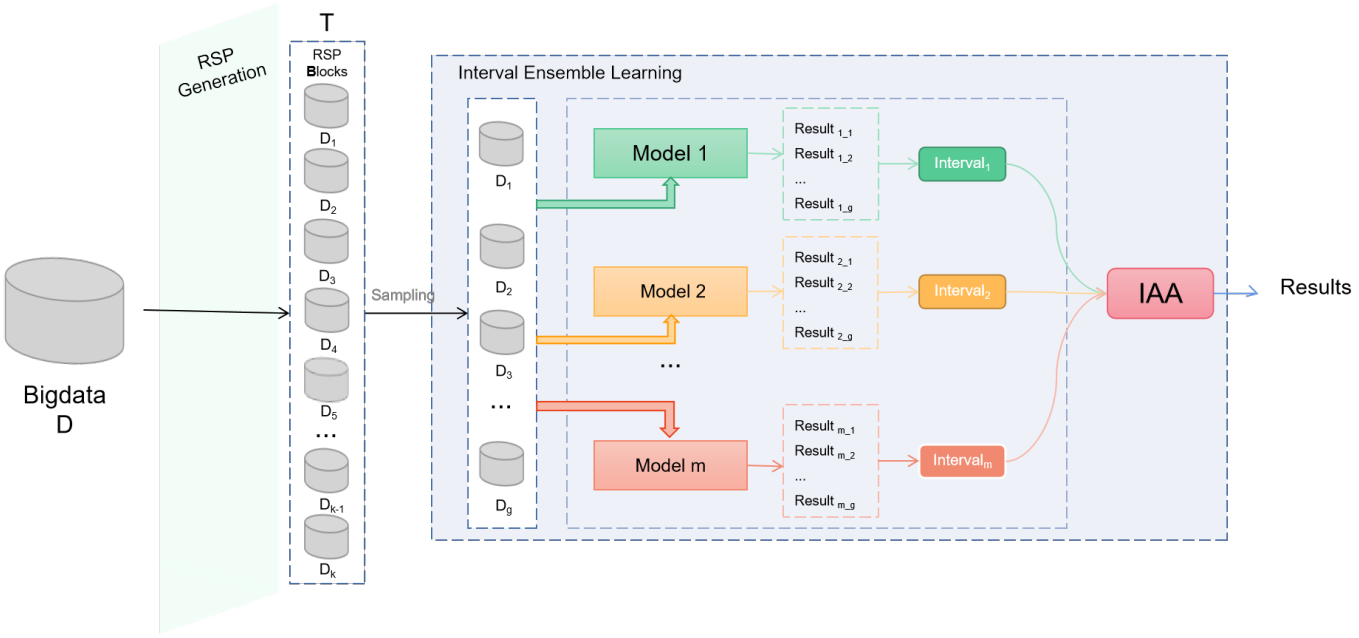


Fig. 2. The Inr-RSP ensemble model for big data analysis

V. Experiments and Results

To demonstrate the classification performance of the proposed model for big data analysis, we conducted several experiments on four datasets. First, we show the characteristics of datasets, experiment settings, and evaluation methods used in our experiments. Then, we evaluate the performance of the proposed model in classification, compared with the traditional RSP analysis model and interval RSP model which are applicable to majority voting. Also, we run our model on different sampling sizes, the various number of selected RSP blocks and learning models to obtain the sensitivity of the proposed model.

A. Datasets

We evaluate the proposed model on four datasets from the University of CaliforniaIrvine (UCI) [1] machine learning repository. As the optimization problem is an ensemble learning under big data, the number of records in the selected datasets is relatively large. In general, each of the four datasets differs in size, features and classes. The properties are described in Table I.

TABLE I
Properties of the datasets used in experiments

Dataset	Records(N)	Features(M)	Classes
Covertype	581,012	54	7
Watch_acc	3,777,046	5	18
SUSY	5,000,000	18	2
HIGGS	11,000,000	28	2

B. Experiment settings

The experiments focus on the classification task, so the decision tree is used as the base classifier. To generate different classifiers, with M as the number of features in each data, each decision tree is generated by abandoning a random feature that has not been ignored. Therefore, the maximum number of classifiers m cannot exceed the number of data features.

Notably, the classifiers' outputs in Inr-RSP are main class probabilities, not labels. For binary classification, the specified primary class is class Zero, and the secondary is class One. Also, the proposed model is suitable for multiclass classification. Consider the class of maximum probabilities as the primary class and the second maximum as the secondary. The multiclass classification ensemble problem is converted to determine the main classification class.

We use the abbreviations below for simplicity. RSP is to represent the traditional RSP analysis model which processes the RSP blocks independently and then aggregates them by majority voting method. Then, using interval RSP to present interval modeling of RSP blocks and aggregation by majority voting. Finally, Inr-RSP is proposed by this paper to represent the interval modeling of RSP blocks but aggregation using the IAA.

In Inr-RSP, the maximum number of classifiers is equal to the number of data features. In the preliminary experiment, to facilitate fair comparisons with other models, the size

[1] https://archive.ics.uci.edu/ml/index.php

of each RSP block n, the number of RSP blocks(g=5) and classifiers(m=5) are fixed for each dataset. In the parameter influence experiment, each dataset is divided into two RSP block sizes. The number of RSP blocks varies from 2 to 20 with intervals of 2, and the number of classifiers differs according to the data features. Each experiment only changes one parameter to reflect the influence of the parameter. To eliminate chance, the experiments are repeated 10 times and the average results are reported.

C. Evaluation Methods

To get convincing results, we use the same testing data to test models for proposed and compared models. Also, using the following two matrices, Accuracy and Kappa, to measure the performance of classification tasks.

$$\text{Accuracy } = \frac{TP + TN}{TP + FN + TN + FP} \tag{7}$$

where TP is the number of true positive predictions, TN is the number of true negative predictions, FP is the number of false positive predictions, and FN is the number of false negative predictions. It shows the proportion of accurate results among the total number of testing.

$$\text{Kappa} = \frac{P_o - P_e}{1 - P_e} \tag{8}$$

where P_o is the overall accuracy and P_e is the chance consistency error. It represents the percentage of errors reduced if the classification were completely random.

D. Preliminary Results

Table II shows the four classification matrices results of the proposed Inr-RSP model compared with the interval RSP model and the RSP model both aggregated by majority voting method in four datasets. For fair comparisons, we used the same sampling size, the number of selected RSP blocks, and the number of learning models for each model. That is, the difference in the results is the performance improvement.

As seen in Table II, both interval ensemble models are superior to the RSP model in most datasets, which means that interval modeling can combine multiple predictions into more accurate interval predictions to preserve the uncertainty. Also, the Inr-RSP ensemble model using the IAA aggregation method successfully outperforms the contrast model in all datasets because the IAA algorithm can well consider the impact of interval values and process uncertain data, which can also transform interval values into more stable and reliable results, so as to make more accurate and robust classification and minimize information loss.

E. Influence of Parameters

In this section, we experiment with the sensitivity of Inr-RSP to changes in parameters, including the RSP sampling size, the number of selected RSP blocks, and the number of classifiers. Note that in evaluating the selected parameters, all other parameters remain fixed during the experimental run.

1) RSP sampling size n:

Figures 3 and 4 present the classification accuracy of the Inr-RSP model on four datasets for two different RSP sampling sizes n (shown as solid lines). As n affects the amount of data, too small n will not allow the classifier to capture enough specific patterns, and so large n may increase the risk of overfitting. It is observed that the RSP sampling size affects the accuracy of the Inr-RSP model, a larger value of n generally leads to better classification.

2) Number of RSP blocks g:

Figure 3 also shows the classification accuracy of the Inr-RSP model on four datasets for different RSP block numbers. Since the traditional RSP model has the same number of classifiers as the blocks, it is not compared without fixed m. As shown in Figure 3, the classification accuracy of the proposed Inr-RSP model increases with the number of RSP blocks at a fixed m = 5 and maintains convergence at a certain number of blocks, indicating that a stable model can be built with a few blocks. In contrast, most of the Inr-RSP models aggregated by the majority voting method are unstable and have poor accuracy as it does not consider the uncertainty of interval data.

3) Number of classifiers m:

Figure 4 represents the classification accuracy of the Inr-RSP model on four datasets for distinct classifier numbers. Because the number of classifiers in the traditional RSP model is the same as the number of blocks, only display the results of g=5 and m=5. The results show that the accuracy of the Inr-RSP model does not have a significant effect so fewer classifiers can build a stable model but no limit to the number of blocks. Meanwhile, it outperforms the comparison model when the g increases.

VI. Conclusion

This paper presents a novel ensemble model for big data named Inr-RSP, which better captures the uncertainty of the traditional RSP-based ensemble model through interval modeling and interval aggregation methods. The Inr-RSP model uses the IAA aggregation method to transform the interval-valued data generated by RSP data blocks interval modeling into fuzzy sets and then obtains the final result through centroid calculation. This model can reduce information loss and obtain more accurate and robust ensemble results. The new model outperforms the traditional RSP model on four real datasets and is also superior to the majority voting method using the IAA.

References

[1] Bo-Wei Chen, Wen Ji, and Seungmin Rho. Divide-and-conquer signal processing, feature extraction, and machine learning for big data. *Neurocomputing*, 174:383, 2016.

[2] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, 2008.

TABLE II
CLASSIFICATION ACCURACY AND KAPPA RESULTS ON FOUR DATASETS

Dataset	n	Accuracy			Kappa		
		RSP	Interval RSP	Inr-RSP	RSP	Interval RSP	Inr-RSP
Covertype	44150	0.8377	0.8375	**0.8619**	0.7372	0.7369	**0.7761**
Watch_acc	35800	0.6605	0.6970	**0.7323**	0.6411	0.6801	**0.7166**
SUSY	50000	0.7444	0.7736	**0.7930**	0.4835	0.5397	**0.5787**
HIGGS	10890	0.6757	0.6836	**0.7109**	0.3499	0.3657	**0.4193**

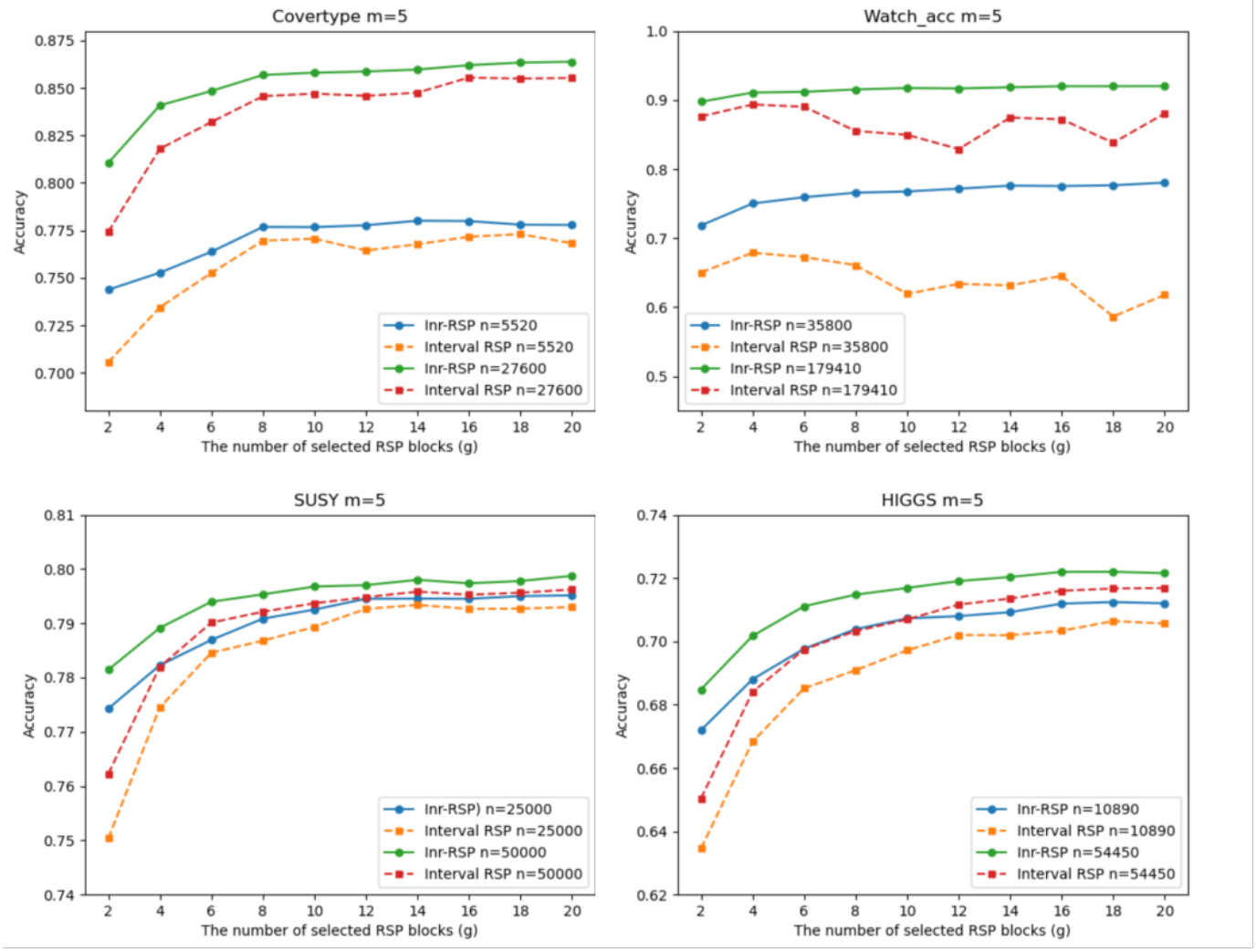


Fig. 3. Compare classification accuracy for the increasing number of RSP blocks with two RSP sampling sizes

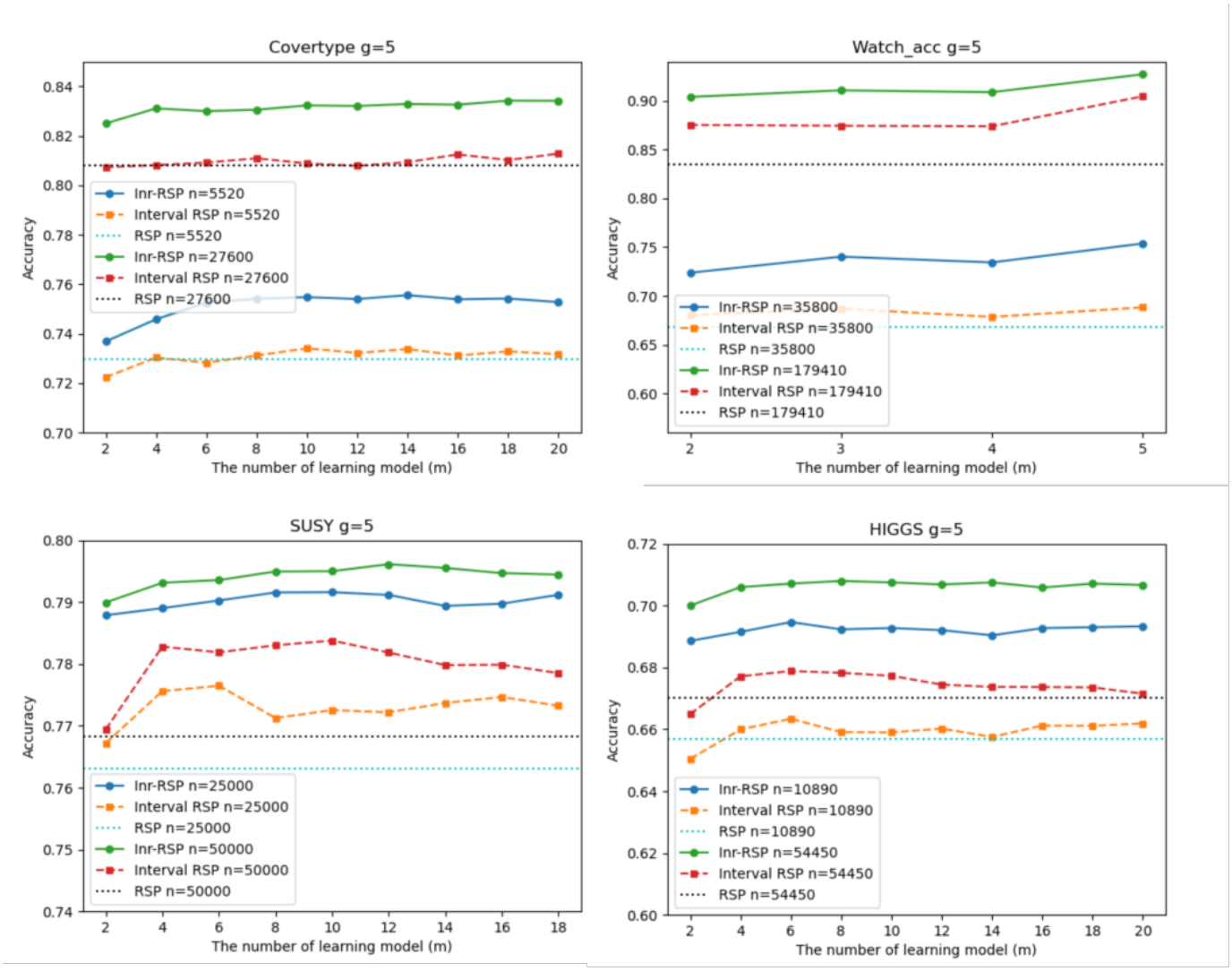


Fig. 4. Compare classification accuracy for the increasing number of classifiers with two RSP sampling sizes

[3] Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica. Spark: Cluster computing with working sets. In Erich M. Nahum and Dongyan Xu, editors, *2nd USENIX Workshop on Hot Topics in Cloud Computing, HotCloud'10, Boston, MA, USA, June 22, 2010*. USENIX Association, 2010.

[4] Lei Gu and Huan Li. Memory or time: Performance evaluation for iterative operation on hadoop and spark. In *10th IEEE International Conference on High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing, HPCC/EUC 2013, Zhangjiajie, China, November 13-15, 2013*, pages 721–727. IEEE, 2013.

[5] Cem Kadilar and Hulya Cingi. Ratio estimators in simple random sampling. *Appl. Math. Comput.*, 151(3):893–902, 2004.

[6] Peter J Bickel and David A Freedman. Asymptotic normality and the bootstrap in stratified sampling. *The annals of statistics*, pages 470–482, 1984.

[7] Jeffrey Scott Vitter. Random sampling with a reservoir. *ACM Trans. Math. Softw.*, 11(1):37–57, 1985.

[8] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler. The hadoop distributed file system. In *2010 IEEE 26th symposium on mass storage systems and technologies (MSST)*, pages 1–10. Ieee, 2010.

[9] Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. A survey on ensemble learning. *Frontiers Comput. Sci.*, 14(2):241–258, 2020.

[10] Leo Breiman. Bagging predictors. *Mach. Learn.*, 24(2):123–140, 1996.

[11] Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In Lorenza Saitta, editor, *Machine Learning, Proceedings of the Thirteenth International Conference (ICML '96), Bari, Italy, July 3-6, 1996*, pages 148–156. Morgan Kaufmann, 1996.

[12] Salman Salloum, Joshua Zhexue Huang, and Yulin He. Random sample partition: A distributed data model for big data analysis. *IEEE Trans. Ind. Informatics*, 15(11):5846–5854, 2019.

[13] Chenghao Wei, Salman Salloum, Tamer Z. Emara, Xiaoliang Zhang, Joshua Zhexue Huang, and Yu-Lin He. A two-stage data processing algorithm to generate random sample partitions for big data analysis. In Min Luo and Liang-Jie Zhang, editors, *Cloud Computing - CLOUD 2018 - 11th International Conference, Held as Part of the Services Conference Federation, SCF 2018, Seattle, WA, USA, June 25-30, 2018, Proceedings*, volume 10967 of *Lecture Notes in Computer Science*, pages 347–364. Springer, 2018.

[14] Aytug Onan, Serdar Korukoglu, and Hasan Bulut. A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. *Expert Syst. Appl.*, 62:1–16, 2016.

[15] CT Fan, Mervin E Muller, and Ivan Rezucha. Development of sampling plans by using sequential (item by item) selection techniques and digital computers. *Journal of the American Statistical Association*,

57(298):387–402, 1962.
[16] Zhicheng Liu and Aoqian Zhang. A survey on sampling and profiling over big data (technical report). *CoRR*, abs/2005.05079, 2020.
[17] Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, and Michael Jordan. The big data bootstrap. *arXiv preprint arXiv:1206.6415*, 2012.
[18] Ludmila I Kuncheva and Juan J Rodríguez. A weighted voting framework for classifiers ensembles. *Knowledge and information systems*, 38:259–275, 2014.
[19] David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.
[20] Adib Ashfaq A Zamil, Sajib Hasan, Showmik MD Jannatul Baki, Jawad MD Adam, and Isra Zaman. Emotion detection from speech signals using voting mechanism on classified frames. In *2019 international conference on robotics, electrical and signal processing techniques (ICREST)*, pages 281–285. IEEE, 2019.
[21] Rahma Atallah and Amjed Al-Mousa. Heart disease detection using machine learning majority voting ensemble method. In *2019 2nd international conference on new trends in computing sciences (ictcs)*, pages 1–6. IEEE, 2019.
[22] Zhiyong Lv, Tongfei Liu, Cheng Shi, Jon Atli Benediktsson, and Hejuan Du. Novel land cover change detection method based on k-means clustering and adaptive majority voting using bitemporal remote sensing images. *Ieee Access*, 7:34425–34437, 2019.
[23] Mikel Uriz, Daniel Paternain, Iris Dominguez-Catena, Humberto Bustince, and Mikel Galar. Unsupervised fuzzy measure learning for classifier ensembles from coalitions performance. *IEEE Access*, 8:52288–52305, 2020.
[24] Christian Wagner, Simon Miller, Jonathan M. Garibaldi, Derek T. Anderson, and Timothy C. Havens. From interval-valued data to general type-2 fuzzy sets. *IEEE Trans. Fuzzy Syst.*, 23(2):248–269, 2015.
[25] Urszula Bentkowska and Barbara Pekala. Diverse classes of interval-valued aggregation functions in medical diagnosis support. In Jesús Medina, Manuel Ojeda-Aciego, José Luis Verdegay Galdeano, Irina Perfilieva, Bernadette Bouchon-Meunier, and Ronald R. Yager, editors, *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Applications - 17th International Conference, IPMU 2018, Cádiz, Spain, June 11-15, 2018, Proceedings, Part III*, volume 855 of *Communications in Computer and Information Science*, pages 391–403. Springer, 2018.
[26] Krzysztof Dyczkowski. *Intelligent Medical Decision Support System Based on Imperfect Information - The Case of Ovarian Tumor Diagnosis*, volume 735 of *Studies in Computational Intelligence*. Springer, 2018.
[27] Urszula Bentkowska, Jan G. Bazan, Wojciech Rzasa, and Lech Zareba. Application of interval-valued aggregation to optimization problem of k-nn classifiers for missing values case. *Inf. Sci.*, 486:434–449, 2019.
[28] Feilong Liu and Jerry M. Mendel. Encoding words into interval type-2 fuzzy sets using an interval approach. *IEEE Trans. Fuzzy Syst.*, 16(6):1503–1521, 2008.
[29] Simon Coupland, Jerry M. Mendel, and Dongrui Wu. Enhanced interval approach for encoding words into interval type-2 fuzzy sets and convergence of the word fous. In *FUZZ-IEEE 2010, IEEE International Conference on Fuzzy Systems, Barcelona, Spain, 18-23 July, 2010, Proceedings*, pages 1–8. IEEE, 2010.
[30] Christian Wagner and Hani Hagras. zslices - towards bridging the gap between interval and general type-2 fuzzy logic. In *FUZZ-IEEE 2008, IEEE International Conference on Fuzzy Systems, Hong Kong, China, 1-6 June, 2008, Proceedings*, pages 489–497. IEEE, 2008.

Cross-Knowledge Graph Relation Completion for Non-isomorphic Cross-lingual Entity Alignment

Yuhong Zhang[1,2], Dan Lu[1,3], Chenyang Bu[1,3], Kui Yu[1,3], Xindong Wu[1,3]
[1]Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology), Hefei 230009, China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230088, Anhui, China
[3]School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China
Corresponding: Yuhong Zhang(zhangyh@hfut.edu.cn)

***Abstract*—The Cross-Lingual Entity Alignment (CLEA) aims to find the aligned entities that refer to the same identity from two Knowledge Graphs (KGs) in different languages. In real-world applications, the neighborhood structures of the same entities in different KGs tend to be non-isomorphic, which makes the entity representation contain diverse semantics information and poses a great challenge for CLEA. In this paper, we address this challenge from two perspectives. On the one hand, cross-KG relation completion rules are designed with the alignment constraint of entities and relations to improve the isomorphism of two KGs. On the other hand, a representation method combining isomorphic weights is designed to include more isomorphic semantics for counterpart entities, which will benefit CLEA. Experimental results show that our model can improve the isomorphism of two KGs and the alignment performance, especially for two non-isomorphic KGs.**

Keywords: Knowledge Graphs, Cross-Lingual Entity Alignment, Non-isomorphism, Relation Completion

I. INTRODUCTION

Knowledge Graphs (KGs) play an important role in NLP field and data mining-related fields, such as question answering [4], industrial and academic settings [12]. But the construction of KGs is very hard that needs substantial resources. Due to the scarcity of available resources, it is difficult to build KGs for under-resourced languages, such as Greek, Arabic, etc. To address this problem, recent research has proposed Cross-Lingual Entity Alignment (CLEA) to enhance KG of under-resourced language using well-resourced language [4].

CLEA is to identify the aligned entity pairs referring to the same objects from two KGs in different languages. To this end, CLEA methods try to map the entities and relations in two KGs into a shared space, in which, the embeddings of the same objects in two KGs are as close as possible. Existing CLEA methods are classified into TransE-based methods [11] and Graph Neural Network-based (GNN-based) methods [22]. TransE-based methods assume that two KGs in different languages have a similar structure, so the embeddings of aligned entity pairs should have relative similar positions in the vector spaces. Recently, GNN-based methods have gained a lot of attention due to their great performance. GNN-based methods first learn the entity embeddings by aggregating the neighboring entities and then evaluate the similarity between entities based on their embeddings. The entities with the nearest geometric distance are regarded as a pair of aligned entities. These methods have proven their effectiveness for the isomorphic KGs.

However, owing to imbalanced resources and different cultures, two KGs in different languages are non-isomorphic generally. Particularly, the ratio of non-isomorphic neighbors is more than 85% for two KGs [17]. The non-isomorphism means that the counterpart entities in two KGs tend to contain heterogeneous neighboring entities, the different numbers of neighbors and relations. As shown in Fig. 1, Given two non-isomorphic KGs and some aligned entities as supervised seeds, (represented by the same shape in yellow), we aim to find more new aligned pairs, such as “林肯” and “Lincon” (red double dashed line). However, The neighborhoods of “林肯” and that of “Lincon” are heterogeneous. They have different neighbors. However, GNN-based methods aggregate these heterogeneous neighbors, which will lead to monolingual embeddings containing different semantics. Thus, the non-isomorphism will hold back CLEA and pose a great challenge.

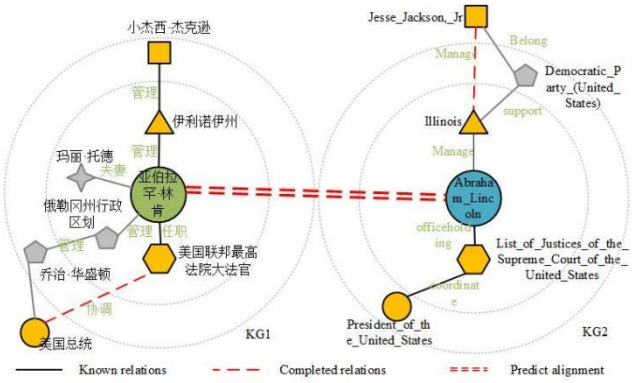


Figure 1. The illustration of non-isomorphism of KGs and our idea of cross-KG relation completion. The neighborhoods of “林肯” and “Lincon” are heterogeneous. Our idea is to change the topology of two neighborhoods by completing the relations (red dashed lines).

A few studies have focused on non-isomorphic CLEA. Their common idea is to expand the neighboring scopes or filter noisy neighbors. Alinet [17] thinks that distant neighbors may include more homogeneous entities and expands neighborhoods to cover more neighbors. DAEA [15] identifies the useful neighborhood with the importance of relations, and then the embedding will include similar neighbors and exclude noisy ones. However, the above methods try to find available information from a single KG, which has two limitations.

DOI reference number: 10.18293/DMSVIVA23-051

1) Although expanding the neighbors' scope can cover more homogeneous neighbors, it inevitably covers some noisy neighboring entities. Moreover, with the increasing scope, the cost of representation methods also increases exponentially. In addition, the expansion of neighbor scope cannot improve the topology isomorphism of two KGs.

2) Existing methods enrich the representation by paying more attention for closer or more similar neighbors. If these closer neighbors are non-isomorphic, the representation of entities will focus more on heterogeneous neighbors and lead to semantic difference.

To address these problems, we propose a method named cross-KGs relation completion for non-isomorphic CLEA. Our method addresses the non-isomorphism in two views. Firstly, with the assumption that counterpart entities should have isomorphic neighborhoods, cross-KGs relation completion is designed to change the topology of KGs and improve the isomorphism of two KGs, as shown with the red dashed lines in Fig.1. Secondly, the isomorphic weights are introduced into representation learning to make entity representation focus more on the isomorphic neighbors, which will benefit non-isomorphic CLEA. Our contributions are summarized as:

1) To address the non-isomorphic CLEA, we propose to improve the topology isomorphism of KGs by cross-KG relation completion. To our best knowledge, there is little work focusing on cross-KG completion owing to the unavailable connection KGs [25]. In this paper, we complete the relations with some supervised aligned information. With our cross-KG completion, both completeness and topology isomorphism can be improved. And KG representation will cover more isomorphic information on a smaller neighborhood.

2) To reduce the semantic discrepancy of counterpart entities, the isomorphic weights for two neighborhoods, not the similarity or importance to the central entity, are introduced into representation learning. The isomorphic weights will make the embedding include more isomorphic semantics and exclude non-isomorphic semantics, making CLEA more easily.

II. Related Work

A. Methods for CLEA

Existing CLEA methods are classified into TransE-based methods [11] and Graph Neural Network-based (GNN-based) methods [22]. And recent studies have shown that GNN-based methods can achieve outperformance. Gnn-based methods can be divided into two types: 1) **GNN with entity attention.** As an expansion of GNN, Graph Convolutional Network (GCN) can learn the node-level representation. GCNAlign [22] is the first study using GCN to learn the representation in low-dimensional space, and then measures the distance of entities to find new alignments. Some works [5] put two KGs into one GCN to learn a shared representation space, in which, it uses aligned pairs to make entities closer to each other. To represent the entities with more semantics, Graph Attention network (GAT) is used to make the representation focus more on the important or similar neighbors [20]. 2) **GCN with relation attention.** To measure the importance of neighbors accurately, some studies [13] combine the relations with neighbors to find useful neighbors. By giving more attention to those useful neighbors, the representation is enhanced further. Other studies use specific relation to update the attention for neighbors, such as the node attributes [14] and relation types [19][24].

In sum, GNN-based methods have proven their superiority for CLEA. However, they only achieve a good performance for similar knowledge graphs [17].

B. Methods for Non-isomorphic CLEA

Non-isomorphism is common in applications. That means the counterpart entities have non-isomorphic neighbors. It is a huge challenge for CLEA. The studies focus on non-isomorphic CLEA can be divided into two categories. 1) **Using additional information**. KDCoE [10] uses both entity description and multilingual literal description as additional information to co-train the embeddings of entities. N-gram [18] uses the attribute triples to generate the embeddings for attribute characters. Other works [1][21] also merge additional configuration information for entities by entities' attributes. 2) **Changing the range of neighborhood**. AliNet [17] is the first work for non-isomorphic CLEA. It expands the scope to cover more distant neighbors to increase the overlapping KGs. And it uses attention to reduce the noisy neighbors and emphasis the useful neighbors. KE-GCN [23] selects the right relations and their corresponding neighbors from all neighborhoods using translated method [11]. DAEA [15] uses the relation and level attention to filter useless and distant neighbors respectively.

Non-isomorphism has attracted much attention. However, existing methods find available information in one KG while neglecting the information from cross-KGs.

III. Our Proposed Method

Formally, KG is defined as *KG=(E, R, T)*. It consists of a set of entities E and a set of relations R, and the knowledge facts are stored in a collection T in form of triples (h, r, t), where h, t $\in E$ and r $\in R$. Given two non-isomorphic KGs, denoted as *KG1=(E1, R1,T1)* and *KG2= (E2,R2,T2)*, the task of our CLEA is to find new aligned entity pairs using some supervised entity pairs $EP=(e1_i, e2_j) \mid (e1_i \in E1; e2_j \in E2)$.

The framework of our method is shown in Fig.2, which includes three steps. The first step is cross-KG relation completion. With the aligned entities as supervised information, the relations alignment is treated as a constraint to predict the potential relations. This step changes the topology of each KG and then improves the topology isomorphism of KGs. The second step is augmented representation. Isomorphic weights are introduced into GAT to make KG representation focus more on the isomorphic neighbors and then the semantic discrepancy of counterpart entities is reduced. The third step is the alignment with a loss function. The distance is used to search for the nearest entity in the whole space and then to find more alignment pairs.

A. Cross-KGs completion

In this paper, non-isomorphic CLEA is addressed differently. We introduce the cross-KG relation completion to change the structure of original KGs and make them more isomorphic.

It can be easily accepted that the same object in different KGs should have homogeneous neighborhoods. If they do not, there may be some relations missing. Motivated by this, we propose to complete these missing relations according to the aligned entities and relations. The supervised aligned entities are known as seeds, and then we need to align the relations. Owing to that the number of relation types is smaller than that of entities. Thus the alignment of relations is easier. We construct the set of aligned relations $RP=(r1_i, r2_j) \mid (r1_i \in R1; r2_j \in R2)$ as a constraint to the completion, which will reduce the noise in the cross-KG relation completion.

The cross-KG completion searches the potential relations in global KGs, and it includes 2 steps, relation alignment and relation completion. We will describe the two steps in detail.

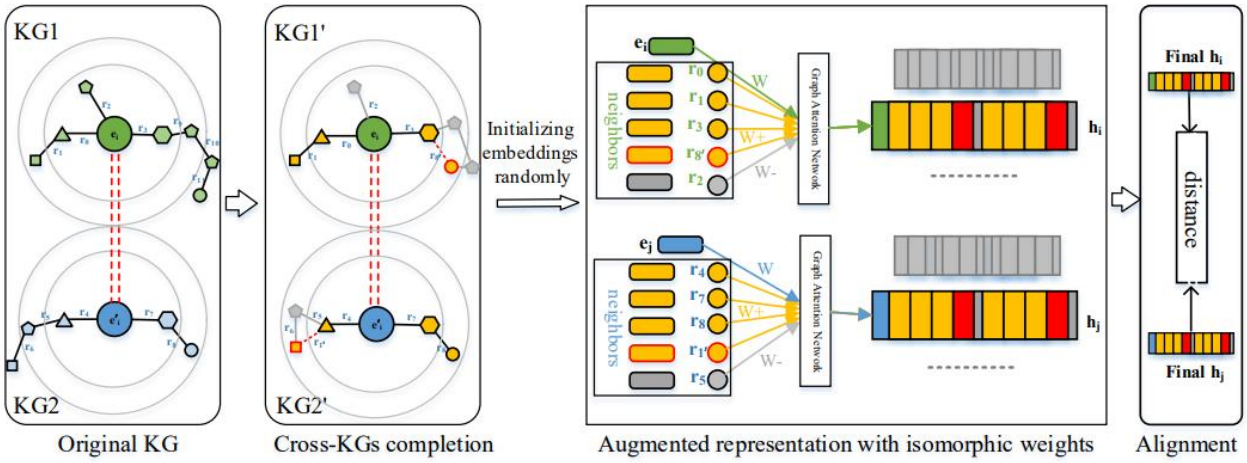


Figure 2. The framework of our method. In cross-KG relation completion step, isomorphic entities in yellow are treated as supervised pairs, and the red dashed lines are the completed relations. In augmented representation step, the yellow W^+ and gray W^- mean isomorphic and non-isomorphic weights respectively.

Relation Alignment Rule: Relations can be aligned based on whether their related entities are aligned. For two triples *(h1, r1, t1)* and *(h2, r2, t2)* from *KG1* and *KG2*, respectively, if the entities *<h1,h2>* and *<t1,t2>* are both aligned, we can infer that *r1* and *r2* should be aligned. Based on this observation, we design the first rule for relation alignment, which is formally expressed as follows.

IF ((h1, r1, t1) ∈ KG1 and (h2, r2, t2) ∈ KG2 and < h1, h2 > ∈ EP and < t1, t2 > ∈ EP)
THEN RP + =< r1, r2 >

Relation Completion Rule: Relations can be completed when two entities are connected in one KG but their aligned entities are not connected in another KG. With the aligned pairs *<h1,h2>*, *<t1,t2>*, and *<r1,r2>* as constraints, if *(h1,r1,t1)* exists in *KG1* but *(h2,r2,t2)* is not included in *KG2*, we will complete the triple *(h2,r2,t2)* in *KG2*. We formally write this as the second rule for relation completion:

IF (< h1, h2 > ∈ EP and < t1, t2 > ∈ EP and < r1, r2 > ∈ RP and (h1, r1, t1) ∈ KG1
THEN KG2 + = (h2, r2, t2)

The relation alignment rule is used to align relations in two KGs, and the aligned relations serve as constraints for relation completion. The relation completion rule is used to complete potential relations, and these completed relations provide more triples for relation alignment. The two rules are run iteratively. Finally, with the completed relations, the neighborhoods of entities change, and the non-isomorphism is reduced.

It is worth noting that the relation-aligned constraint is important for relation completion. Firstly, the aligned relations are introduced as additional information besides neighbors, which is helpful for CLEA. Secondly, the aligned relation constraint connects entities and relations into triples as an aligned unit, ensuring that the completed related neighbors are unambiguous and can reduce noise. If the aligned relations are ignored, we can only connect entities but cannot distinguish their neighbors by relation awareness. Therefore, the relation constraint is necessary.

B. The isomorphic weights for augmented representation

Although the isomorphism of KGs has been improved after completion, it cannot ensure the complete isomorphism of KGs. There still are some heterogeneous neighbors. Thus, we propose isomorphic weights to focus more on homogeneous neighbors and ignore heterogeneous ones in representation. And then the entity embedding will be more suitable for non-isomorphic CLEA. In this subsection, we first set different weights for isomorphic neighbors and non-isomorphic neighbors, and then learn the representation KGs. We take *KG1* as an example to show the weighted aggregation representation, and the representation of *KG2* is similar.

Isomorphic Weight Setting: For two counterpart entities, if their neighbors are known as aligned seeds, they are called isomorphic neighbors. It can be defined as follows.

Isomorphic Neighbors: Given $e_i \in KG1$ and $e_j \in KG2$, and $N_{ei} = n_i^{ei}$ and $N_{ej} = n_j^{ej}$ denote the neighbors of e_i and e_j. If $< n_i^{ei}, n_j^{ej} > \in EP$, they are called isomorphic neighbors. Otherwise, they are non-isomorphic neighbors. The isomorphic value is set as $I(n_i^{ei}, n_j^{ej})$. When $I(\cdot) = 1$, it means n_i^{ei} and n_j^{ej} are isomorphic, and when $I(\cdot) = -1$, it means they are heterogeneous.

$$I(n_i^{ei}, n_j^{ej}) = \begin{cases} 1 & < n_i^{ei}, n_j^{ej} > \in EP \\ -1 & < n_i^{ei}, n_j^{ej} > \notin EP \end{cases} \quad (1)$$

The isomorphic weights of neighbors are set according to whether they are isomorphic. Especially, the weights are initialized equally, and then the isomorphic neighbors will make the weight $W_{n_i^{ei}}$ larger owing to $I(\cdot) = 1$, and the non-isomorphic ones will make the weight smaller with $I(\cdot) = -1$. $W_{n_i^{ei}}$, the weight for *i*-th neighbor of entity e_i, is calculated as following.

$$W_{n_i^{ei}} = \frac{W_{initial}}{Z_{|N_{ei}|}} exp(I(n_i^{ei}, n_j^{ej})) \quad (2)$$

where $W_{initial}$ is the initial weight of neighbors, and it is set to one out of the number of neighbors equally. $Z_{N_{ei}}$ is the normalized factor used to normalize the weight value, and it is calculated as:

$$Z_{N_{ei}} = \sum_{i=1}^{|N_{ei}|} W_{initial} exp(I(n_i^{ei}, n_j^{ej})) \quad (3)$$

Weighted Augmented Representation: Isomorphic weights are combined with GAT to learn KG representation. Firstly, embeddings of entities and relations are initialized randomly, denoted as $h_{e_1}, h_{e_2}, \ldots, h_{e_k}$ and $h_{r_1}, h_{r_2}, \ldots, h_{r_k}$. Then the isomorphic weights are set to all neighbors and relations to augment those isomorphic neighbors. The weighted aggregation of relations and neighbors is shown in Formula 4 and 5.

$$h_{e_i,r} = \frac{1}{|N_{ei}|}(\sum_{r_k} W_{n_i^r} h_{r_k}) \quad (4)$$

$$h_{e_i,N} = \frac{1}{|N_{ei}|+1}(\sum_{e_i \in N_{ei}} W_{n_i^{ei}} h_{n_k^{ei}} + W_{initial} h_{ei}) \quad (5)$$

where h_{r_k} refers to the embeddings of relations associated with entity e_i, and $h_{n_k^{ei}}$ refers to the embeddings of neighbors that belong to entity e_i. Combining both relations and neighbors enables the representation of the entity e_i.

$$h_{e_i} = [h_{e_{i,r}} || h_{e_{i,N}}] \quad (6)$$

Secondly, the weighted embeddings h_{ei} are used as the input of GAT to learn the final representation of *KG*. Formula 7 shows the learning representation process of GAT.

$$h_{e_i}^{final} = ReLU(\frac{1}{Z}\sum_{z=1}^{1}[\sum_{e_i \in N_{ei}} \alpha_{i,j}^z h_{ei}]) \quad (7)$$

where Z is the number of head attention, $\alpha_{i,j}^z$ is the attention. Both Z and $\alpha_{i,j}^z$ are computed as MRAEA [8] does.

C. Entity Alignment

With the representation KGs, we find new entity pairs by searching the nearest entity to each other globally [22]. In this process, the distance between entities can be computed by the Manhattan distance.

$$dis(e_1, e_2) = |\hat{h}_{e_1}^{out} - \hat{h}_{e_2}^{out}| \quad (8)$$

where $\hat{h}_{e_1}^{out}$ and $\hat{h}_{e_2}^{out}$ represent the embeddings of e_1 and e_2.

To bring similar entities closer to each other in a uniform space, we shorten the distance by minimizing the following loss.

$$L = \sum_{<e_1,e_2> \epsilon EP} ReLU(dis(e_1, e_2) - dis(e'_1, e_2) - dis(e_1, e'_2) + \lambda \quad (9)$$

where λ represents the margin hyper-parameter. The entities e'_1 and e'_2 are considered as negative entities. We randomly select negative pairs from *E1* and *E2*, similar to MRAEA [8]. As shown in Formula 9, our calculations enhance positive samples and weaken negative samples in order to narrow the alignment entity distance.

IV. Experiment

A. Datasets and Baselines

The performance of our method is evaluated on three large cross-lingual datasets from DBP15K, which are used commonly in many studies. For this dataset, the ratio of overlap coefficient (OC) is used to show the isomorphism of KGs. The OC is proposed in [17], and it is computed by the ratio of aligned neighbors to all neighborhoods in one-hop neighboring range. Higher the OC value, the more isomorphic two KGs [17]. For example, the OC value of ZH-EN is 11.7%, which means that only 11.7% of neighborhoods in Chinese and English KGs are homogeneous. It can be seen that the two KGs are non-isomorphic. the OC value of JA-EN is 11.6% and the OC value of FR-EN is 13.1%. Baselines

To validate the effectiveness of our method, fifteen baselines are compared with our method. These baselines fall into 3 categories. TransE-based baselines include MTransE [11], IPTransE [26], and NAEA [27]. GNN-based baselines include GCN-Align [22], MuGCN [2], GAT [20], R-GCN [13], MuGCN [2], MRAEA [8], RREA [9], Dual-AMN [7] , PSR [6], Sparse[3] and RpAlign [16]. There are also some baselines focusing on the non-isomorphic CLEA, including of AliNet [17], KE-GCN [23] and DAEA [15].

It is worthy to note that a few recent works [1][3][28] have achieved remarkable performances. [1][28] use some additional information, such as entities' attribute information and entities' description information. [3][28] initialize the representation with Glove embedding. In this paper, we compare our methods with its variants igoring the additional information for fair comparision.

B. Experimental Setting

DBP15K consists of three cross-lingual tasks, namely $DBP_{ZH\text{-}EN}$, $DBP_{JA\text{-}EN}$, and $DBP_{ER\text{-}EN}$. For a fair comparison, we use 30% of alignments data as training and the other 70% as testing, as other methods did [8]. In addition, there are some common parameters, which all are set to the same values as the previous works. The embeddings' dimensions of entities and relations *d=100*, attention head number *k=2*, the depth of GNN is set to 2, the dropout rate is 0.3 and the learning rate of Adam

is 0.005. The margin-based loss function integrates some negative entities. The aggregation range, dropout rate, and learning ratio are set to 2, 0.3, and 0.005 respectively. In this paper, Hits@k and Mean Reciprocal Rank (MRR) are used to measure performance.

C. Main Results

Table I shows the performance of our method and baselines. Experimental results of all baselines are obtained from their original papers. Some conclusions can be drawn from Table I.

1) GCN-based methods outperform TransE-based methods, which is consistent with the conclusion of other works [19].

2) As for GCN-based methods, methods with GAT [7], [20] perform better than those with GCN. It is because that the similar or closer entities are given more attention to enrich the representation of KGs. In addition, the GNN-based methods focusing on both relations and entities [9], [13] perform better than those only focusing on the entities.

3) The methods for non-isomorphic CLEA, including AliNet [17], KE-GCN [23], DAEA [15] and ours, perform better than the GCN-based methods on average owing to that they address the non-isomorphism of KGs. It shows that the non-isomorphism does exist commonly in two KGs and addressing it will benefit the CLEA.

4) Our method has an obvious improvement than other non-isomorphism baselines, including AliNet [17], KE-GCN [23], DAEA [15], KE-GCN [23] and RpAlign[16]. H@1 of our method is improved by 14.5%, 13.2%, and 14.9% averagely on three datasets. Compared with AliNet [17], our method not only covers more neighbors but also changes the topology of the neighbors using cross-KG relation completion. By improving the isomorphism of KGs, our method achieves an improvement. And compared with KE-GCN, our method enriches the entity embedding by supplementing missing homogeneous neighbors rather than deleting heterogeneous neighbors, which includes more related and similar semantic information. Compared with RpAlign[16], Our completion rule depends on non-isomorphic relation and assign isomorphic weights to make the representation include more isomorphic information.

TABLE I. OVERALL PERFORMANCE OF ALL METHODS ON DBP15K DATASET

Types	Models	DBP_{ZH-EN}			DBP_{JA-EN}			DBP_{FR-EN}		
		H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR
TransE-based CLEA	**MTransE [11]**	30.8	61.4	36.4	27.9	57.5	34.9	24.4	55.6	33.5
	IPTransE [26]	40.6	73.5	51.6	36.7	69.3	47.4	33.3	68.5	45.1
	NAEA [27]	65.0	86.7	72.0	64.1	87.2	71.8	67.3	89.4	75.2
GCN-based CLEA	**GCN-Align [22]**	41.3	74.4	54.9	39.9	74.5	54.6	37.3	74.5	53.2
	GAT [20]	41.8	66.7	50.8	44.6	69.5	53.7	44.2	73.1	54.6
	R-GCN [13]	46.3	73.4	56.4	47.1	75.4	57.1	46.9	75.8	57.0
	MuGCN [2]	49.4	84.4	61.1	50.1	85.7	62.1	49.5	87.0	62.1
	MRAEA [8]	65.7	89.5	74.4	72.7	92.3	79.8	73.9	93.8	81.0
	RREA [9]	71.5	**92.9**	79.0	71.3	**93.3**	79.3	73.9	94.6	81.6
	Dual-AMN [7]	73.1	92.3	79.9	72.6	92.7	79.9	75.6	**94.8**	**82.7**
	PSR [6]	70.2	92.4	78.1	69.8	93.0	78.2	73.1	94.1	80.7
	Sparse [3] (L=0)	58.5	78.0	-	59.1	79.1	-	76.0	91.5	-
non-isomorphism CLEA	**AliNet [17]**	53.9	82.6	62.8	54.9	83.1	64.5	55.2	85.2	65.7
	DAEA [15]	56.8	88.3	67.7	57.6	89.2	68.3	58.0	91.2	69.5
	KE-GCN [23]	56.2	84.2	66.4	57.0	85.2	67.0	57.2	88.5	68.3
	RpAlign [16]	74.7	88.8	79.4	72.9	89.0	78.2	75.2	89.9	80.1
	Ours	**74.9**	92.2	**80.6**	**73.8**	92.5	**83.0**	**76.3**	93.4	81.7

TABLE II. ABLATION OF OUR METHOD ON DBP15K DATASET

Models	DBP_{ZH-EN}			DBP_{JA-EN}			DBP_{FR-EN}		
	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR
Baseline(AliNet)	53.9	82.6	62.8	54.9	83.1	64.5	55.2	85.2	65.7
Baseline(MRAEA)	65.7	89.5	74.4	72.7	92.3	79.8	73.9	93.8	81.0
W/O isomorphic weights	72.9	90.5	79.7	73.3	92.3	81.8	74.4	93.0	81.2
W/O rel Completion	73.8	91.5	83.0	73.1	92.1	79.9	74.5	93.1	81.3
Ours	74.9	92.2	80.6	73.8	92.5	83.0	76.3	93.4	81.7

D. Ablation Studies

Ablation is conducted in Table II. w/o rel completion means ignoring the cross-KG relation completion and finds new aligned pairs from the original KGs. And w/o isomorphic weights means ignoring isomorphic weights and learns the representation only with GAT.

The effectiveness of cross-KGs relation completion: Compared with MRAEA, w/o isomorphic weights improves H@1 by 2.7% average. Compared with w/o rel completion, our method also improves H@1. It reveals that cross-KG completion can improve the completeness and isomorphism of KGs.

The effectiveness of isomorphic weights: Compared with w/o isomorphic weights, our method improves H@1 by 3.56% average. This reveals the augmented representation for isomorphic neighborhoods can enhance KG representation and is helpful for non-isomorphic CLEA.

E. Analysis

1) ***Cross-KG relation completion can improve the completeness and isomorphism of KGs.***

Cross-KG relation completion results are shown in Table III.

a) After completion, the number of triples increases by 32455 and 6173 for KG_{ZH} and KG_{EN} respectively. It means 32455 and 6173 relations are completed and the completeness of KGs is improved. With the increase of the number of isomorphic edges, the isomorphism of the graph is enhanced, so that the representation of aligned entities is closer, and the training results of the graph neural network are more accurate.

b) After completion, OC values are improved by 8.8%, 9.3%, and 9.5%, which shows that isomorphism of KGs is improved.

TABLE III. THE RESULT OF CROSS-KG RELATION COMPLETION

Datasets	Indicators	Original	After Completed	Increase
DBP_{ZH-EN}	$Triples_{ZH}$	70,414	102,869	32,455
	$Triples_{EN}$	95,142	101,317	6,175
	OC	11.7%	20.5%	8.8%
	H@1	67.0%	74.9%	7.9%
DBP_{JA-EN}	$Triples_{JA}$	77,214	89,804	12,590
	$Triples_{EN}$	93,484	120,489	27,005
	OC	11.6%	20.9%	9.3%
	H@1	55.2%	73.8%	18.6%
DBP_{FR-EN}	$Triples_{FR}$	105,998	206,658	100,660
	$Triples_{EN}$	115,722	155,477	39,755
	OC	13.1%	22.6%	9.5%
	H@1	55.2%	76.3%	21.1%

c) With the improvement of completeness and isomorphism, H@1 is improved by 7.9%, 18.6%, and 21.1%. This shows the effectiveness of the cross-KG relation completion.

2) Our method can achieve an identical performance only covering the least neighboring scopes.

Some baseline methods, such as AliNet [17], expand the neighborhood scope to improve the isomorphism between two KGs. We compared the performance of AliNet and our method with varying ranges, as shown in Fig.Figure 3. .

a) When the neighborhood range changes from 1-hop to 2-hops, the performance of AliNet improves significantly, indicating that expanding the scope covers more homogeneous neighbors. However, when the scope changes to 3-hops and 4-hops, the performance of AliNet decreases sharply, demonstrating that a larger scope covers more heterogeneous and noisy neighbors, which hinders CLEA.

b) As the neighborhood scope increases, the performance of both our method and our w/o weight remains relatively stable. This is because our method changes the topology of all entities through cross-KG completion. After completion, the isomorphism between two KGs will not change when the neighborhood scope varies. Additionally, as the neighborhood scope expands, the isomorphic weights weaken distant and heterogeneous neighbors. This implies that our method does not require aggregating too many neighbors in the representation learning.

3) ***The robustness for non-isomorphism of KGs.***

Figure 4. shows the performance of our method with different OC. The larger the OC value is, the stronger isomorphism of knowledge graph is. We randomly drop out some homogeneous neighbors from the original KGs to get several datasets with different OC. We delete [5%-30%] isomorphic neighbors and OC value will decrease from 11.1% to 8.1% for ZH-EN, 11.0% to 8.1% for JA-EN and 12.4% to 9.2% for FR-EN.

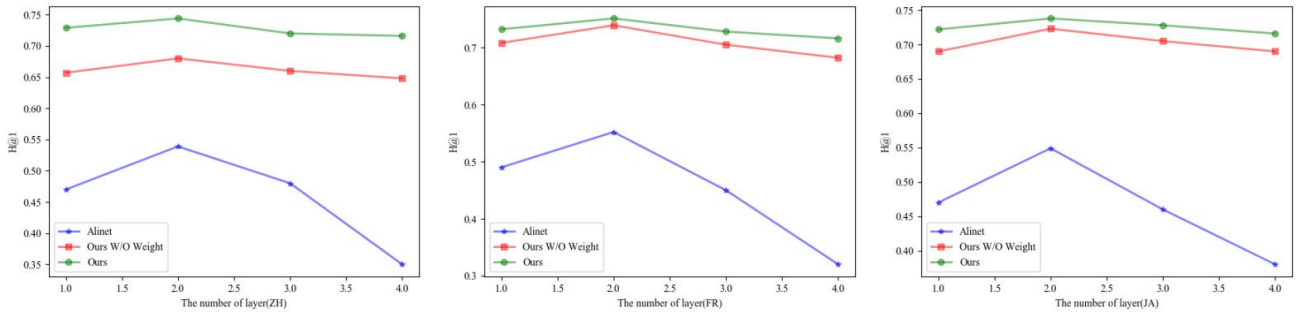


Figure 3. The influence for performances of different ranges of neighborhoods on ZH-EN, JA-EN, FR-EN.

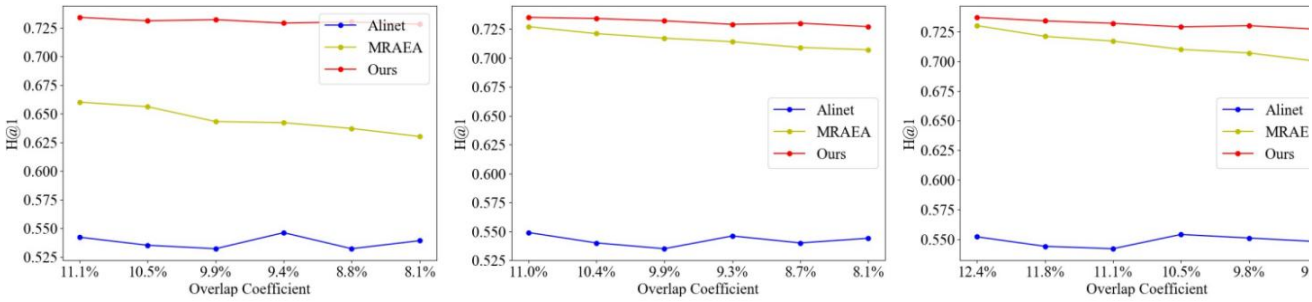


Figure 4. The performance varies with different OC values on three datasets.

a) With the decreasing of OC, the performances of our method and MRAEA [8] decrease obviously. It shows that non-isomorphism will hold back CLEA. While the performance of Alinet fluctuates over a range because it randomly selects distant neighbors and does not rely on direct neighbors.

b) Compared with MRAEA[8], our method degrades more slowly, and performs relatively higher and more stably. This reveals that our method is robust, especially for non-isomorphic CLEA.

4) ***The robustness for available alignment seeds.***

All baselines in this paper are supervised methods. The number of available seeds will influence the alignment performance. In our method, the available seeds will influence both the supervised learning and the cross-KG relation completion. The results of Hits@1 and the OC value varying with the size of aligned entity pairs are shown in Figure 5. and Figure 6.

In Figure 5, the alignment accuracy increases with the number of pre-aligned seed entities increasing. In Figure 6, OC value increases more obviously with the increasing of the seeds number. For Zh-EN, when there are only 1500 seeds available, the OC increases by 1.84%, while 4500 seeds are available, the OC increases by nearly 9%. It shows that more aligned seeds will be conducive to our entity alignment.

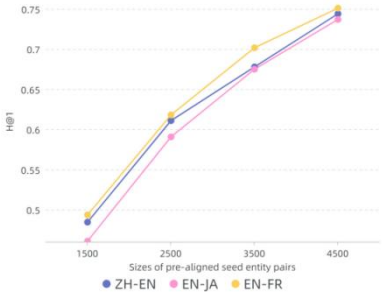


Figure 5. The performance changes with the size of available aligned entity pairs.

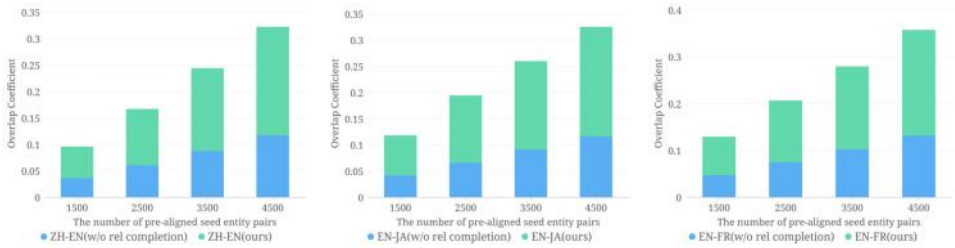


Figure 6. The OC changes with the size of available aligned entity pairs.

V. CONCLUSIONS

This paper focuses on non-isomorphic CLEA. To address the non-isomorphism, cross-KG relation completion is proposed to complete the missing relations and improve the completeness and isomorphism. And then, the isomorphic weights, not the importance of central entities in one KG, are designed to learn a representation more suitable for CLEA. In near future, we will explore more suitable method to mearue the isomorphism of two KGs.

ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Program of China (under grant 2020AAA0106100), the National Natural Science Foundation of China (under grant 61976077), the Natural Science Foundation of Anhui Province (under grant 2208085MF170) and the University Synergy Innovation Program of Anhui Province (GXXT-2022-040).

References

[1] W. Cai, Y. Wang, S. Mao, J. Zhan, and Y. Jiang, "Multi-heterogeneous neighborhood-aware for knowledge graphs alignment." Info. Proc. & Mana., vol. 59, pp. 529-551, 2022.

[2] Y. Cao, Z. Liu, C. Li, Z. Liu, J. Li, and T. Chua, "Multi-channel graph neural network for entity alignment." Proc. 57th Conf. Asso. Comp. Ling., Florence, Italy, vol. 1. Long Papers, pp. 1452–1461, 2019.

[3] M. Fey, J. E. Lenssen, C. Morris, J. Masci, and N. M. Kriege, "Deep graph matching consensus." 8th Inte. Conf. Lear. Repr. Ethiopia, April 26-30, 2020.

[4] H. Cui, T. Peng, F. Xiao, J. Han, R. Han,L. Liu, "Incorporating anticipation embedding into reinforcement learning framework for multi-hop knowledge graph question answering." Info. Scie., vol. 619, pp. 745-76, 2023.

[5] X. Kun, W. Liwei, Y. Mo, F. Yansong, S. Yan, W. Zhiguo, and Y. Dong, "Cross-lingual knowledge graph alignment via graph matching neural network." Proc. 57th Conf. Asso. Comp. Ling., Florence, Italy, vol. 1. Long Papers, pp. 3156–3161, July 28- August 2, 2019.

[6] X. Mao, W. Wang, Y. Wu, and M. Lan, "Are negative samples necessary in entity alignment?: An approach with high performance, scalability and robustness." 30th ACM Inte. Conf. Info. Know. Mana., Queensland, Australia, pp. 1263‑1273, November 1 - 5, 2021,.

[7] X. Mao, W. Wang, Y. Wu, and M. Lan, "Boosting the speed of entity alignment 10 ×: Dual attention matching network with normalized hard sample mining." WWW' 21: Web Conf., Ljubljana, Slovenia, pp. 821‑832, April 19-23, 2021.

[8] X. Mao, W. Wang, H. Xu, M. Lan, and Y. Wu, "Mraea: an efficient and robust entity alignment approach for crosslingual knowledge graph." Proc. 13th Inte. Conf. Web Sear. Data Mini., pp. 420‑428, 2020.

[9] X. Mao, W. Wang, H. Xu, Y. Wu, and M. Lan, "Relational reflection entity alignment." CIKM' 20: 29th ACM Inte. Conf. on Info. Know. Mana., Ireland, pp. 1095‑1104, October 19-23, 2020.

[10] C. Muhao, T. Yingtao, C. Kai-Wei, S. Steven, and Z. Carlo, "Co-training embeddings of knowledge graphs and entity descriptions for cross-lingual entity alignment." Proc. Twen. Inte. Join. Conf. Arti. Inte., Stockholm, Sweden, pp. 3998‑4004, July 13-19, 2018.

[11] C. Muhao, T. Yingtao, Y. Mohan, and Z. Carlo, "Multilingual knowledge graph embeddings for cross-lingual knowledge alignment." Proc. 26th Inte. Join. Conf. Arti. Inte., Melbourne, Australia, pp. 1511‑1517, August 19-25, 2017.

[12] A. Rossi, D. Barbosa, D. Firmani, A. Matinata, and P. Merialdo, "Knowledge graph embedding for link prediction: A comparative analysis." ACM Tran. Know. Disc. Data, vol. 15, pp. 1‑49, 2021.

[13] M. Schlichtkrull, T. N Kipf, P. Bloem, R. V. D. Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks." Euro. Sema. Web Conf., pp. 593‑607, 2018.

[14] C. Shang, Y. Tang, J. Huang, J. Bi, X. He, and B. Zhou. "End-to-end structure-aware convolutional networks for knowledge base completion." Proc. AAAI Conf. Arti. Inte., pp. 3060‑3067, 2019.

[15] J. Sun, Y. Zhou, and C. Zong, "Dual attention network for cross-lingual entity alignment." Proc. 28th Inte. Conf. Comp. Ling., pp. 3190‑3201, 2020.

[16] H. Huang, C. Li, X. Peng, L. He, S. Guo, H. Peng, et al, "Cross-knowledge-graph entity alignment via relation prediction." Know. Syst. vol. 240, pp. 107813, 2022.

[17] Z. Sun, C. Wang, W. Hu, M. Chen, J. Dai, W. Zhang, and Y. Qu, "Knowledge graph alignment network with gated multi-hop neighborhood aggregation." Proc. AAAI Conf. Arti. Inte., pp. 222‑229, 2020.

[18] B. D. Trisedya, J. Qi, and R. Zhang, "Entity alignment between knowledge graphs using attribute embeddings." Proc. AAAI Conf. Arti. Inte., pp. 297‑304, 2019.

[19] S. Vashishth, S. Sanyal, V. Nitin, and P. Talukdar, "Composition-based multi-relational graph convolutional networks." 8th Inte. Conf. Lear. Repr., Addis Ababa, Ethiopia, April 26-30, 2020.

[20] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Li`o, and Y. Bengio, "Graph attention networks." 6th Inte. Conf. Lear. Repr., Vancouver, BC, Canada, April 30 - May 3, 2018.

[21] C. Wang, B. Han, S. Pan, J. Jiang, G. Niu, and G. Long, "Cross-graph: Robust and unsupervised embedding for attributed graphs with corrupted structure." 2020 IEEE Inte. Conf. Data Mini., pp. 571‑580, 2020.

[22] Z. Wang, Q. Lv, X. Lan, and Y. Zhang, "Crosslingual knowledge graph alignment via graph convolutional networks." Proc. 2018 Conf. Empi. Meth. Natu. Lang. Proc., pp. 349‑357, 2018.

[23] D. Yu, Y. Yang, R. Zhang, and Y. Wu, "Knowledge embedding based graph convolutional network." Proc. Web Conf., pp. 1619‑1628, 2021.

[24] Z. Zhang, F. Zhuang, H. Zhu, Z. Shi, H. Xiong, and Q. He, "Relational graph neural network with hierarchical attention for knowledge graph completion." Proc. AAAI Conf. Arti. Inte., pp. 9612‑9619, 2020.

[25] Y. Zheng, "Methodologies for cross-domain data fusion: An overview." IEEE Tran. Big Data, 1(1):16‑34, 2015.

[26] H. Zhu, R. Xie, Z. Liu, and M. Sun, "Iterative entity alignment via knowledge embeddings." Proc. Inte. Join. Conf. Arti. Inte., pp. 4258‑4264, 2017.

[27] Q. Zhu, X. Zhou, J. Wu, J. Tan, and L. Guo, "Neighborhood-aware attentional representation for multilingual knowledge graphs." Proc. 28th Inte. Join. Conf. Arti. Inte., Macao, China, pp. 1943‑1949, August 10-16, 2019.

[28] R. Zhu, M. Ma, and P. Wang, "RAGA: relation-aware graph attention networks for global entity alignment." Adva. Know. Disc. Data Mini. 25th Paci. Conf., pp. 501‑513, May 11-14, 2021

Directional Residual Frame: Turns the motion information into a static RGB frame

Pengfei Qiu, Yang Zou*, Xiaoqin Zeng, Xiaoxiang Lu, Xiangchen Wu
Institute of Intelligence Science and Technology, School of Computer and Information,
Hohai University, Nanjing, China
{qiupf0718, yzou, xzeng}@hhu.edu.cn

Abstract*—The most commonly adopted methods of video action recognition are optical flow and 3D convolution. Optical flow method requires calculation in advance and a lot of computing resources. 3D convolution method encounters several problems such as many parameters, difficult training, and redundant computation. This paper proposes an approach that can turn the motion information into a static RGB frame by a feasible way of compression: Directional Residual Frame (DRF). This idea comes from a static cartoon that can represent complex events through residual shadows. DRF takes advantage of the scarce nature of residual frames in space and pixel value to achieve similar effects of residual shadows by fusing multiple residual frames. With the DRF, the motion information can be learnt as simply and efficiently as learning the RGB information. In addition, it also proposes a Short-term Residual Shadow Module based on the DRF. Experimental results show that it has better performance than the state-of-the-art model TDN on UCF101 benchmark.

Keywords-DRF; motion information; action recognition; temporal difference module

I. INTRODUCTION

In recent years, Video-based action recognition has drawn a significant amount of attention from the academic community. In action recognition, there are two kinds of key and complementary information: appearances and motion. CNN have achieved great success in classifying images of objects, scenes, and complex events. Thus, it is crucial for action recognition to capture motion information in video, which is usually achieved by two kinds of mechanisms in the current deep learning approaches: two-stream network [1] and 3D convolutions [5,6,7]. Even though the two-stream network can effectively improve the accuracy of action recognition through the optical flow, it requires a lot of computing resources to extract the optical flow. Although the 3D convolution can learn motion features directly from the RGB frames, it also leads to large network models and high computational cost.

Therefore, how to efficiently learn motion information has been a crucial challenge in action recognition.

In everyday life, we can know the motion information of the meteor, the fan and other things through the residual shadow. Obviously, we acquire the motion information from a certain moment of spatial information. Think about it the other way. Can we use a 2D frame to characterize a complex movement process? The optical flow can only reflect the speed, and requires multiple pieces to characterize the non-linear motion. Comics are a very successful case in point. A cartoon can represent a wonderful fight by using a residual shadow. The shadow in static cartoons can be well characterized in the complex motion process. And temporal derivative (difference) is highly relevant to optical flow [2], and has shown effectiveness in action recognition by using RGB difference as an approximate motion representation [3, 4].

In this paper, we propose a motion representation approach based on RGB difference, termed as Directional Residual Frame (DRF). The principle of DRF is similar to the shadow in comics that turns the motion information into a static RGB frame. First, we subtract two adjacent frames in the video with absolute value to obtain residual frames [11]. Then, the residual frames are binarized. During the binarization process, the motion features are retained and the difference caused by the brightness change is removed. Finally, the adjacent residual frames are fused to form a residual shadow-like trajectory map. As shown in Figure 1, our DRF is a good representation of the trees moving to the right (the movement caused by the camera movement) and the people running to the left.

Figure 1. The first 5 frames are consecutive frames in the video, and the sixth frame is the corresponding DRF.

To demonstrate the effectiveness of the DRF, we performed the experimental analysis using Temporal Difference Network (TDN) [12] on the benchmark UCF101 [13], which is the state-of-the-art method without optical flow and 3D convolution. TDN is able to yield a state-of-the-art performance on both motion relevant Something-Something V1 datasets [9] and

*Corresponding author: yzou@hhu.edu.cn(Y. Zou)

DOI reference number: 10.18293/DMSVIVA23-007

scene relevant Kinetics datasets [10], under the setting of using similar backbones[12].

The technical contributions of the paper are summarized as follows:

1) To reduce the serious redundant calculation in video understanding, we propose an effective compression approach DRF that can turn the motion information into a static RGB information by using the scarcity of residual frame, due to the high similarity between adjacent frames. Optical flow requires multiple stacks to react non-linear motion, whereas DRF demands only one.

2) Based on the DRF, we propose a Short-term Residual Shadow Module (S-RSM) to capture the motion information.

3) The experimental results show that compared with the S-TDM in the state-of-the-art model TDN, our approach achieves higher accuracy with fewer model parameters.

The rest of the paper is organized as follows. Section 2 proposes the concept and calculation process of the DRF, and presents the S-RSM module based on the DRF; Section 3 describes the details of the experiments and evaluates the effectiveness of our method on UCF101 benchmark; and Section 4 concludes the paper.

II. DIRECTIONAL RESIDUAL FRAME

In this section, we describe the proposed DRF in detail. First, we give an overview of DRF. Then, we elaborate the calculation process of the DRF. Finally, we provide the implementation details of using DRF in TDN.

A. Overview

Residual shadow is the most successful case that turns RGB information into the motion information. Residual shadow has both motion trajectory information and direction information. So how to form a residual shadow from continuous frames is a challenge. Our thoughts of addressing this include two steps, as follows:

First, motion detection. In this step, the motion information is extracted from the sequential RGB frame. Objects undergoing spatial position changes in the image sequence are presented as foreground (motion region).

Second, motion fusion. In this step, the motion information extracted from the previous step is fused into a static RGB frame where the motion region blurs with time, like residual shadow.

Motion detection. The common methods for motion detection are: background subtraction, temporal difference and optical flow [17]. Both background subtraction and optical flow require a lot of computing resources, which are contrary to efficiency. Therefore, we adopt the temporal difference method to extract the motion object. The temporal difference method may mistakenly detect the area originally covered by the object as moving, called Ghost, which is a problem with motion detection. As shown in Figure 2, the area originally covered by the moving object will be incorrectly detected into motion, which is the Ghost. But Ghost will not be a problem here, because it can be used effectively in motion fusion.

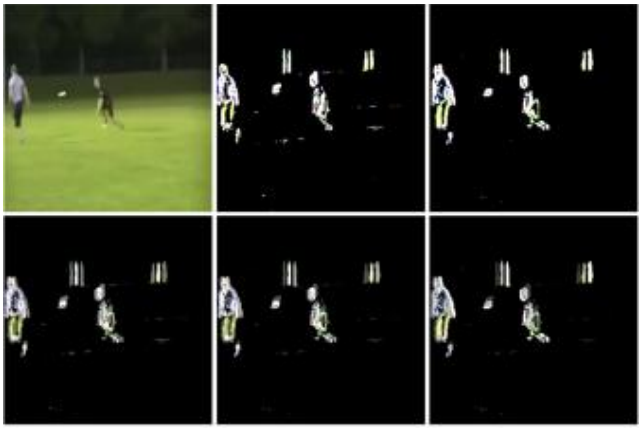


Figure 2. Motion Detection. The second and third frames are the two consecutive temporal differences before the first RGB frame. The fourth frame is (df1 + df2), the fifth is (2 * df2 + df1), and the sixth is the DRF, where df1 is Frame 2, and df2 is Frame 3.

Motion fusion. Motion fusion is the core of the proposed approach.

How to represent the direction of the movement is a crucial issue. In Figure 2, although the fourth and fifth frame preserve more motion traces with the scarcity of the residual frame, there is not any temporal information (direction information). The direction of motion is recognized with the aid of the numerical growth direction. In DRF, objects move from dark to light. From the sixth frame in Figure 2, it can be easily seen through residual shadow that the trees are moving to the right.

$$\begin{bmatrix}0&1&0&0\\1&0&1&0\\0&1&0&0\\0&0&0&1\end{bmatrix}+2*\begin{bmatrix}0&1&0&1\\1&0&0&1\\0&0&0&0\\0&0&0&1\end{bmatrix}+2^2*\begin{bmatrix}1&0&0&1\\1&0&1&0\\0&1&0&0\\0&0&0&0\end{bmatrix}=\begin{bmatrix}4&3&0&6\\7&0&5&2\\0&5&0&0\\0&0&0&3\end{bmatrix}$$

Figure 3. Binary fusion. Every matrix of 5 * 5 represents a residual frame, and the region of the matrix with an element value of 1 indicates the foreground.

Another issue is how to preserve the complete information in the motion fusion process. Although the residual frame is scarce, the foreground of different residual frames may overlap. The overlapping region of the foreground of two adjacent residual frames is the Ghost of the latter residual frame. As for more than two frames, the overlapping region will become difficult to interpret. The overcoverage approach, where the overlapping region takes the same value as the late residual frame, would lose a lot of information. Our approach is inspired by the binary coding to use the value-domain scarce nature of the residual frame. Each number of pixel values indicates an overlapping possibility. In Figure 3, the region with an element value of 7 in the resulting matrix is the overlapping region of 3 matrices; and the region with a value of 5 is the overlapping region of the first and third matrices.

B. The calculation process of the DRF

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and

others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

The process of calculating the RDF is divided into three steps. First, Temporal difference is employed to remove background and acquire motion region. Then, the binarization is adopted to remove the noise and obtain the scarce residual frame. Finally, the DRF is obtained from the fusion of residual frames.

Step 1: Residual frames

Residual frames contain more motion-specific features by removing still objects and background information and leaving mainly the changes between frames [11]. As shown in the third frame in Figure 4, the movement region will be brighter than the static areas.

The movement regions in the residual frame achieve positive or negative values, which are highly correlated with the pixel value of the background. The correlation can cause the movement regions being either positive or negative in the residual frame, thus failing to know the direction of the object. In Figure 4, the Frisbee is white with values above the background color, so it moves from the negative to the positive area in the residual frame. But this direction of movement is unreliable. Therefore, we utilize the absolute residual frame to alleviate the interference of the pixel value of the background. The issue of motion direction in the absolute residual frame will be tackled in step 3.

Figure 4. The first three frames are adjacent frames, the fourth one is the corresponding residual frame, the fifth one is the residual frame after the absolute value, and the sixth one is a binarization of the fourth one.

Here we use $Frame_i$ to represent the *ith* frame data, and $Frame_{i\sim j}$ denotes the stacked frames from the *ith* frame to the *jth* frame. The process of obtaining residual frames can be formulated as follows:

$$ResFrame_{i\sim j} = \left| Frame_{i\sim j} - Frame_{i+1\sim j+1} \right| \tag{1}$$

At this stage, the Residual frames is not a sparse matrix. Influenced by the camera motion and light intensity changes, the gray area is not all *0*.

Step 2: Binarization

Binarization of the residual frames: *0* is used to represent no change area, and *1* is used to represent change area. In the field of image segmentation, there are a few algorithms [14] for image binarization. In this paper, we adopt threshold method in order to reduce the amount of computation as much as possible.

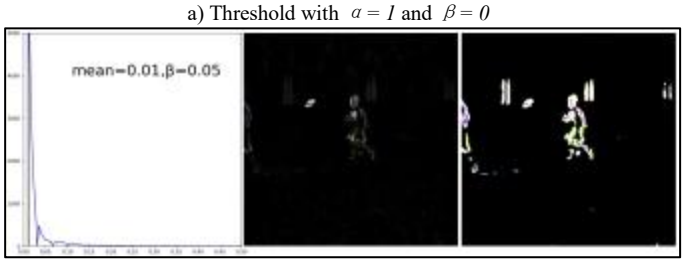
The formula is as follows:

$$threshold = \frac{\alpha}{n^2} * \sum_{x=1}^{n}\sum_{y=1}^{n} ResFrame_i(x,y) + \beta \tag{1}$$

$$ResFrameB_i(x,y) = \begin{cases} 1 & ResFrame_i(x,y) > threshold \\ 0 & otherwise \end{cases} \tag{2}$$

Here $ResFramei(x,y)$ is the image value of the coordinates (x, y) in the *ith* residual frame. α and β are super parameters. And n is the size of the *ith* residual frame.

a) Threshold with $\alpha = 1$ and $\beta = 0$



b) Threshold with $\alpha = 1$ and $\beta = 0.05$

Figure 5. Binarization. The "mean" in the images represents the mean of the residual frames. The first chart of each row is the pixel value statistics chart, the second is the residual frame, and the third is the binarized residual frame.

Since residual frame is a scarcity matrix, the mean value tends to be below the minimum in the movement region. With very small amplitude of motion, the mean may be lower than the difference value caused by changes in light intensity. We denote the minimum of the threshold by β. Figure 5 illustrates the effect of β on removing the background noise.

Step 3: Motion fusion. The higher the value is, the later the event occurs.

The motion fusion of multiple binary residual frames transforms temporal information into numerical information. The higher the value is, the later the event occurs.

$$DRFrame_n^{'} = \sum_{i=1}^{n} 2^{i-1} * ResFrameB_i \tag{3}$$

We accumulate the residual frames according to formula 3. There may be overlaps between the differences of consecutive residual frames. Various overlapping cases of n residual frames will be mapped to the value $0\sim2^n$. The case with $n=4$ is shown in Figure 6. In Figure 3, the region in the result matrix corresponding to the motion region (value is *1*) in the third

matrix should obtain the maximum value to indicate the movement end point. But the value of overlapping region will be greater than the last motion region. The brightest region appears in the middle region of the motion trajectory, as in the fifth frame of Figure 2.

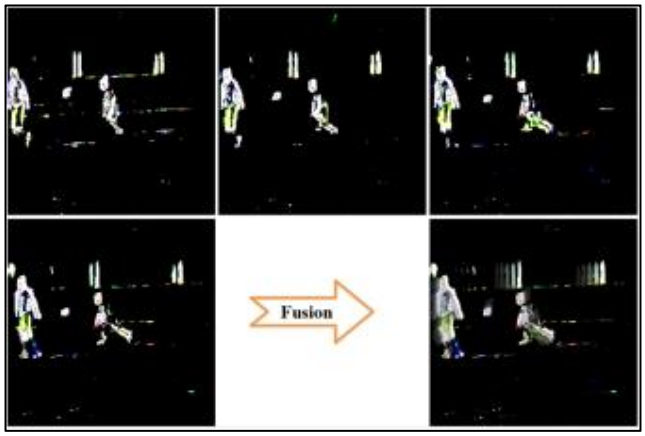


Figure 6. The first four pictures are continuous residual frames after binarization, and the last one is the DRF fused by the first four.

$$Mask_i(x,y)=\begin{cases}1 & DRFrame_n'(x,y)=2^{i-1}\\ 0 & otherwise\end{cases} \quad (4)$$

$$DRFame_n = DRFame_n' + \sum_{i=1}^{n} 2^{i-1}\left(ResFrameB_i * Mask\right) \quad (5)$$

In Formula 4, the operator sets the non-zero element in the matrix to 0 and the zero element to 1. Then, through Formula 5, the value of the non-overlapping difference is doubled.

C. S-RSM with DRF

Based on the DRF, a Short-term Residual Shadow Module (S-RSM) is proposed, as an improvement of the S-TDM in TDN, as illustrated in Figure 8.

Temporal Difference Networks (TDN) is a video-level architecture of capturing both short-term and long-term information for end-to-end action recognition. TDN is composed of a short-term and long-term temporal difference module (TDM), as illustrated in Figure 7 [12]. In Figure 9, the short-term TDM in TDN supply a single RGB frame with a temporal difference to yield an efficient video representation, explicitly encoding both appearance and motion information [12].

In TDN, the stacks of difference frames are processed by 2D convolution, which can only capture limited motion information and of which the main function is to calibrate the moving area on the static image.

The DRF turns the action information into the static RGB information by fusing multiple temporal difference frames. So the model can capture the movement information by learning the RGB information in the DRF. This feature of DRF is beneficial to 2D convolutional networks to learn motion features, so as to perform the task of action recognition even better.

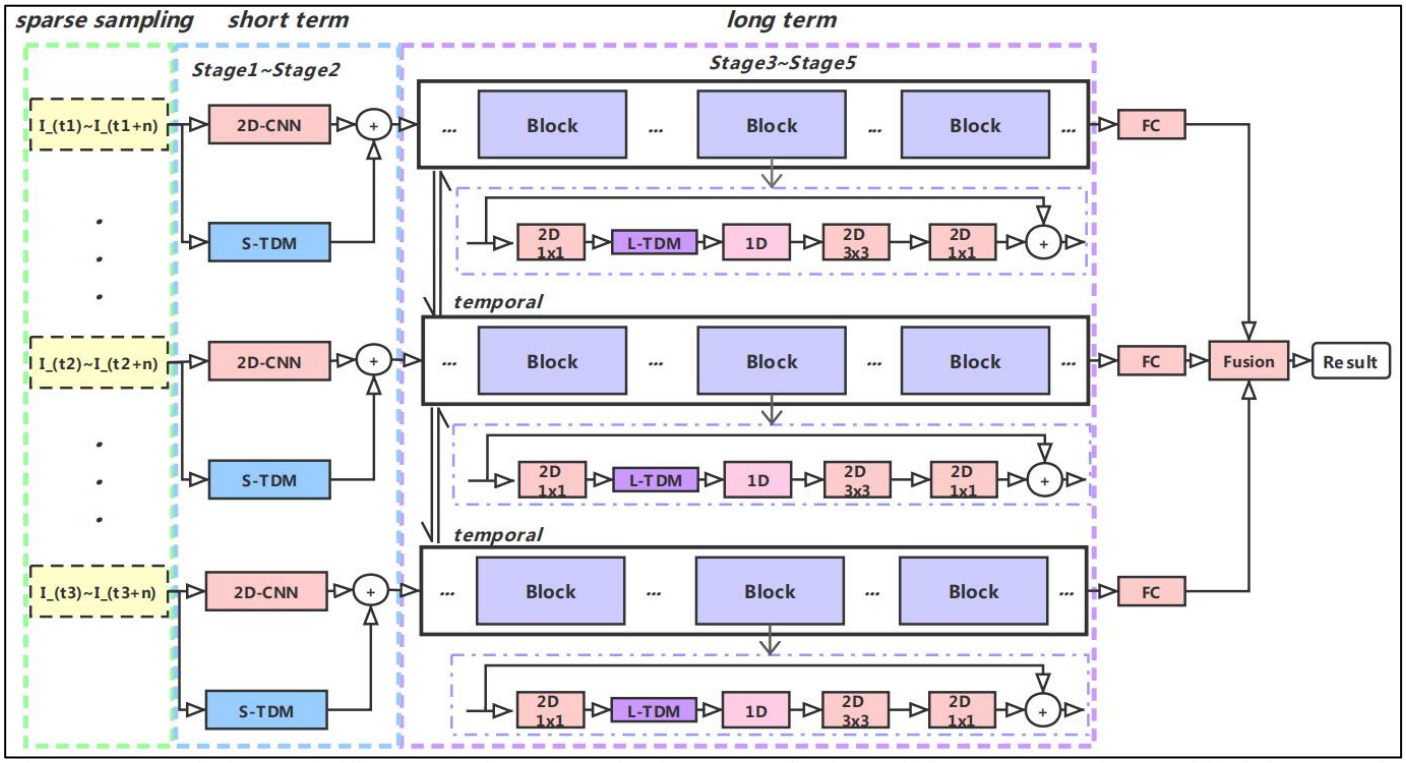


Figure 7. Framework of Temporal Difference Network (TDN). Based on the sparse sampling from multiple segments, TDN aims to model both short-term and long-term motion information.

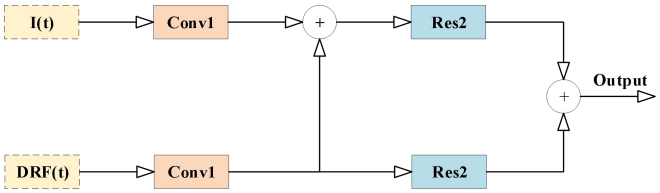


Figure 8. Framework of short-term Residual Shadow Module with DRF

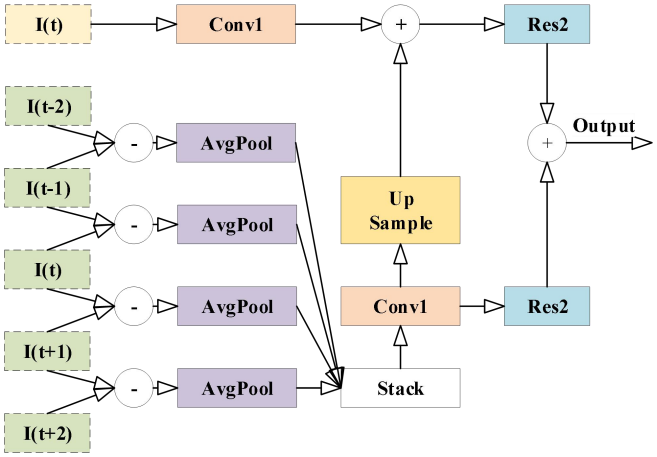


Figure 9. Framework of the short-term TDM.

III. EXPERIMENTS

In this section, we present the experiment results of the proposed DRF. First, we describe the evaluation datasets and implementation details. Then, we evaluated the effectiveness of DRF on the state-of-the-art method TDN.

A. Datasets and implementation details

Video datasets. There are several commonly used datasets for video recognition tasks. We mainly focus on the benchmark: UCF101. UCF101 consist of 13,320 videos in 101 action categories [13].

Training and testing. In experiments, we use ResNet50 to implement TDN framework. Following common practice [15], during training, each video frame is resized to have shorter side in [256, 320] and a crop of 244 × 244 is randomly cropped. We pre-train TDN on the ImageNet dataset [16]. The batch size is 128 and the initial learning rate is 0.02. The total training epoch is set to 60 in the UCF101 benchmark. The learning rate will be divided by a factor of 10 when the performance on validation set saturates. For testing, the shorter side of each video is resized to 256. We implement the kind of testing scheme: 1-clip and center-crop where only 1 center crop of 244 × 244 from a single clip is used for evaluation.

B. Experimental Results

From the experimental results in Table I, it can be found that the *beta* of DRF taking 0.05 is a suitable value. The binarized threshold as the mean has lower accuracy than the other two schemes, which confirms the viewpoint we mentioned in Section 2. With very small amplitude of motion, the mean may be lower than the difference value caused by changes in light intensity.

Since the residual frame is absolute, the *beta* of DRF is the lower limit of the threshold. The accuracy of the *beta* of DRF being 0.05 is higher than that of the *beta* of DRF being 0.1. When the threshold is set too high, some important motion information will be filtered out.

TABLE I. ACC OF DIFFERENT BINARIZATION PARAMETERS ON UCF101 BENCHMARK

Method	Backbone	Input (S-RSM)	Frames	Length (DRF)	Alpha (DRF)	Beta (DRF)	Top-1 (UCF101)
TDN	ResNet50	DRF	3	5	1.00	0.00	84.51%
TDN	ResNet50	DRF	3	5	1.00	0.05	85.33%
TDN	ResNet50	DRF	3	5	1.00	0.10	84.92%

From the experimental results in Table Ⅱ, it can be shown that the frames of the DRF motion fusion is of length 5 in UCF101 benchmark. When the DRF length is set to 3, the reason for the decrease of accuracy is that TDN learns too little action information, whereas it is set to 7 and 9, the reason for the decrease of accuracy is that it is difficult for TDN to learn.

TABLE II. ACC OF DIFFERENT MOTION FUSION LENGTH ON UCF101 BENCHMARK

Method	Backbone	Input (S-RSM)	Frames	Length (DRF)	Alpha (DRF)	Beta (DRF)	Top-1 (UCF101)
TDN	ResNet50	DRF	3	3	1.00	0.05	84.29%
TDN	ResNet50	DRF	3	5	1.00	0.05	85.33%
TDN	ResNet50	DRF	3	7	1.00	0.05	84.48%
TDN	ResNet50	DRF	3	9	1.00	0.05	84.48%

The results in Table Ⅲ show that the proposed TDN outperforms the original model at sampling frames of 4 and 8. With the sample frame of 4, our approach improves by nearly 1% over the original method; and with the sample frame of 8, our approach improves by more than 1.2%.

TABLE III. ACC OF DIFFERENT MODULE ON UCF101 BENCHMARK

Method	Backbone	Input	module	Frames	Pretrain	Top-1 (UCF101)
TDN (original)	ResNet50	RGB + difference	S-TDM	4	ImageNet	84.97%
TDN (original)	ResNet50	RGB + difference	S-TDM	8	ImageNet	87.15%
TDN (ours)	ResNet50	RGB + DRF	S-RSM	4	ImageNet	85.95%
TDN (ours)	ResNet50	RGB + DRF	S-RSM	8	ImageNet	88.39%

From this set of comparative experiments, it can be concluded that DRF contains better motion information than the stacked residual frames. In [12], it has been shown that TDN can reach the state-of-the-art level without the use of optical flow and 3D convolution.

IV. CONCLUSION

To address the problem of serious redundant calculation in video motion recognition, we propose the approach to

squeezing the motion information into a RGB frame. The principle of DRF is similar to the shadow in comics. The shadow in static cartoons can be well characterized in the complex motion process. DRF exploits the scarcity of residual maps to fuse the motion information of multiple residual maps into one spatial frame. In this way, it can learn motion information as it learns about RGB information. Based on the DRF, we propose S-RSM based on 2D convolution to capture motion information. Through comparative experiments, we verified that our approach has better performance than the state-of-the-art model TDN in UCF101 benchmark.

REFERENCES

[1] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in NIPS, 2014, pp. 568-576.

[2] Berthold K.P. Horn and Brian G. Schunck, "Determining optical flow," Artificial Intelligence, vol.17, pp. 185-203, 1981.

[3] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. lin, X. Tang and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," European conference on computer vision, vol. 9912, pp. 20-36, 2016.

[4] Z. Yue, Y. Xiong, and D. Lin, "Recognize Actions by Disentangling Components of Dynamics," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6566-6575.

[5] S. Ji, W. Xu, M. Yang and K. Yu, "3D convolutional neural networks for human action recognition." IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 35, pp. 221-231, 2013.

[6] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489-4497.

[7] K. Hara, H. Kataoka, and Y. Satoh, “Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018, pp. 6546–6555.

[8] J. Lin, C. Gan, and S. Han, "TSM: Temporal Shift Module for Efficient Video Understanding," IEEE/CVF International Conference on Computer Vision (ICCV) IEEE, 2019, pp. 7082-7092.

[9] J. Carreira, and A. Zisserman, "Quo vadis, action recognition? a new model and the Kinetics Dataset," Computer Vision and Pattern Recognition IEEE, 2017, pp. 4724-4733.

[10] R. Goyal, SE. Kahou, V. Michalski, J. Materzynska and R. Memisevic, "The" something something" video database for learning and evaluating visual common sense," Proceedings of the IEEE international conference on computer vision, 2017, pp. 5843-5851.

[11] L. Tao, X. Wang, and T. Yamasaki, "Rethinking motion representation: Residual frames with 3D ConvNets," IEEE Transactions on Image Processing, vol. 30, pp. 9231-9244, 2021.

[12] L. Wang, Z. Tong, B. Ji and G. Wu, "TDN: Temporal Difference Networks for Efficient Action Recognition," Computer Vision and Pattern Recognition IEEE, 2021, pp. 1895-1904.

[13] M. S. Hutchinson and V. N. Gadepally, "Video action understanding," IEEE Access, vol. 9, pp. 134611-134637, 2021.

[14] Otsu, N. "A Threshold Selection Method from Gray-Level Histograms," IEEE Transactions on Systems Man & Cybernetics, vol. 9, pp. 62-66, 2007.

[15] C. Feichtenhofer, H. Fan, J. Malik and K. He, "Slowfast networks for video recognition," Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 6201-6210.

[16] L. J. Li, R. Socher and F. F. Li, "Towards total scene understanding: Classification, annotation and segmentation in an automatic framework," IEEE Conference on Computer Vision & Pattern Recognition IEEE, 2009, pp. 2036-2043.

[17] A. A. Shafie, F. Hafiz and M. H. Ali, "Motion detection techniques using optical flow," World Academy of Science Engineering & Technology, 2009

A Topic Lifecycle Trend Prediction Algorithm on Facebook

Chen Luo [0000−0002−5747−742X]
CyberAray Network Technology Co.,Ltd
Shenzhen, CN
chenluo@usc.edu

Jun Shi
CyberAray Network Technology Co.,Ltd
Shenzhen, CN
jshi@nscslab.net

***Abstract*—Recently, social media has been widely used for people to discuss public opinions and share their views. Internet public opinions have attracted a lot of attention from the government, enterprises and the general public. How to properly analyze, utilize and guide these online opinions is an extremely important issue that the world is currently faced with, and the prediction of topic lifecycle trends is the key to solving this problem. This paper proposes a topic lifecycle trend prediction algorithm based on Facebook data. The algorithm takes into account the similarity between new topics and historical topics in terms of lifecycle curves and the similarity in terms of text content, finds a curve that can best represent the future lifecycle trend of new topics, and then effectively predicts the trend of new topics. It is helpful and meaningful to use this method in the early warnings and predictions of online public opinions and hot topics.**

Keywords-Public Opinions;Facebook; Hot topics; Trend prediction; Similarity; Lifecycle

I. INTRODUCTION

Nowadays, social media has become an indispensable part of people's daily lives. According to the Digital 2020 July Global Statshot report released by We Are Social and Hootsuite [1], we know that there are now 3.96 billion social media users worldwide, accounting for about 51% of the world's total population. Therefore, social media is one of the most important ways to propagate and spread information. In those various social media platforms, there are many hot topics generated every day. These hot topics are significant carriers which contain focused information people pay attention to, and they are directly related to the size of the social influence triggered by the events. It is crucial to make good use of the information in hot topics. On one hand, the government can monitor and analyze hot topics to understand the trend of online public opinions and take corresponding measures in time, which is conducive to maintaining the long-term stability of society. On the other hand, enterprises can understand the needs of users through relevant hot topics to make business plans such as personalized marketing to some users. As a result, putting forward a method that can predict the trend and analyze the lifecycle of topics is of great importance.

Facebook, as the world's most popular social media platform with over 2.6 billion monthly active users, has a plenty of topic data. In this paper, we use posts information from Facebook as our datasets, which include date, time, content and some other useful information. We first extract daily hot topics from Facebook daily posts, and then we use Jaccard Similarity algorithm to calculate the similarities between daily hot topics and posts from other days by comparing their keywords in order to find those posts which are related to hot topics. Based on that, we use the number of relevant posts as the value of y-axis, dates as the value of x-axis to plot the lifecycle curve of each hot topic. Topics with similar lifecycle curves are merged into one cluster, which means all topics under the same cluster have similar trends. After that, we extract a centroid curve for each cluster to stand for the trend of the cluster. When a new topic comes, Dynamic Time Warping (DTW) algorithm is used to compute similarities between curves to find curves from all clusters that are most similar to the curve of new topic. From those topics which are similar in curve, we check if their contents are similar to the new topic as well. Considering similarity both in curve and in content, we can find one curve that best represents the future lifecycle trend of a new topic.

On the basis of description above, this paper proposes a topic lifecycle trend prediction algorithm based on Facebook dataset. There is no doubt that we encountered many difficulties in this process, such as finding posts that are related to hot topics, complicate data cleaning, reducing the time complexity of algorithm operation to increase efficiency and so on. With the efforts of keeping trying and optimizing, we finally propose this method. The main contribution of this paper can be summarized as follows:

- We use the K-Shape algorithm to cluster time series data, making topics with similar lifecycle curves into one cluster, which allows us to effectively observe and analyze the characteristics of different types of topic lifecycle, and make effective trend predictions for new topics that have similar curve characteristics to historical ones.
- We use the DTW algorithm to calculate the similarity between new topics and historical topics on the curve, solving the problem that ordinary Euclidean distance cannot compare the similarity of two unaligned or unequal length sequences.
- Considering the similarity of topics both in lifecycle curve and in text content, we propose a method to predict topic lifecycle trend.

The rest of the paper is discussed as the following order, Chapter II introduces the related work. In Chapter III, we present the topic lifecycle trend prediction method. Chapter IV shows the experiments and evaluation. At last, conclusion and future work is discussed in Chapter V.

* DOI reference number: 10.18293/DMSVIVA23-105

II. Related Work

In 2007, Jiang, Yue [2] made a study showing that early lifecycle data can be used to predict the fault-prone modules in a project. In 2012, Shota Ishikawa [3] designed a system detecting hot topics during a certain period of time and a method was proposed to reduce the variation of posted words related to the same topic, which provides a great contribution to AI services. In the same year, Rong Lu and Qing Yang [4] defined a new concept as trend momentum, which are used to predict the trend of news topics. Juanjuan Zhao [5] developed a model of short-term trend prediction of topics based on Sina Weibo dataset while the accuracy still needs to be improved when the trend of topic changes too frequently in 2014. More recently in 2018, Abuhay [6] used NMF topic modeling method to find topics and implemented ARIMA to forecast the trend of research topics. Chaoyang Chen and Zhitao Wang [7] proposed a correlated neural influence model, which can predict the trending research topics among the research evolution of mutually influenced conferences in the same year. In 2021, Yumei Liu and Shuai Zhang [8] researched the use of blockchain technology in the financial field, utilized various kinds of methods like co-word analysis and bi-cluster algorithm to explore hot topics and predict the future development trend. In 2022, a scientific research topic trend prediction model based on multi-LSTM and Graph Convolutional Network was proposed by Mingying Xu and Junping Du [9]. Compared to other baseline models, its experiment results showed an improvement on the prediction precision.

III. Topic Lifecycle Trend Prediction Method

Our goal is to build a topic lifecycle trend prediction algorithm model. Before that there are some necessary work to be done first, including hot topic extraction, finding topic-related posts, drawing historical topic lifecycle graphs, clustering and curve fitting based on lifecycle curve shapes. After all these tasks are completed, we calculate the shape similarity between topic lifecycle curves by using DTW algorithm. In the meantime, we calculate text content similarity based on keyword matching. Considering both shape similarity and text content similarity, we propose a topic lifecycle prediction method to predict the future lifecycle of new topics. The detailed flowchart is shown in Fig 1.

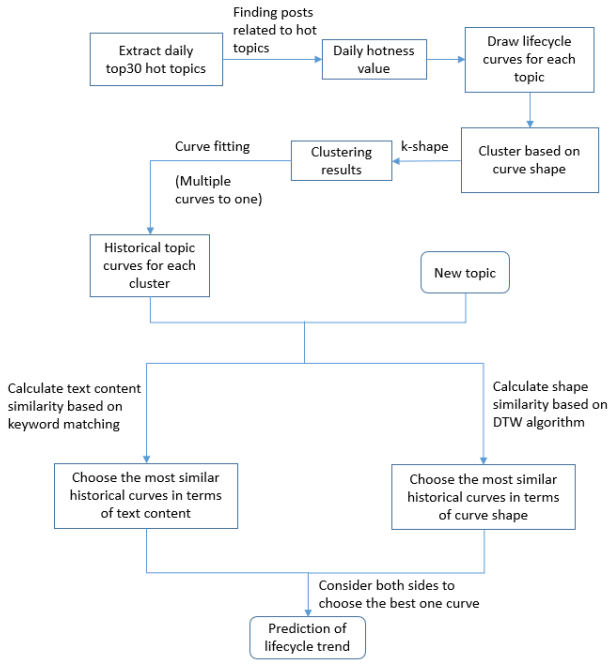


Figure 1. Flow chart of topic trend prediction method.

A. *Hot Topic Extraction*

In this paper, the data we used came from crawling the Facebook platform. We crawled some of the posts' information of the Facebook platform from May 2020 to April 2021 as needed, a total of 100,535,793 posts with non-empty content.

Since there are too many post contents in different languages, we cannot analyze all of them. Then we intend to study the posts' information of only two languages this time, English and Chinese, by considering the number of language users, language popularity and some other factors. Thus, the first step is to identify Chinese and English posts, and then we perform tokenization. For Chinese lexical analysis, we use the tool called HanLP, and for foreign language lexical analysis, we use spaCy. Both HanLP and spaCy are commonly used natural language processing tools. Subsequently, keyword extraction is performed. A combination of lexical analysis, entity recognition, TF-IDF [10] and TextRank [11] algorithms are used to extract keywords, which can be classified by category as people, places, organizations, etc. Then the single pass algorithm is used to cluster the sentences in the post content based on keyword similarity, and the sentences with similar keywords are clustered into one class, that is, into one topic. Next, we need to give each topic a *topic description* to stand for the content of this topic. By achieving that, we extract the keywords of this topic, calculate the similarity between the keywords of the topic and the keywords of each sentence in the topic in order, and select the sentence with the highest similarity score and the shortest length as the topic description of this topic. Besides, we need to know each topic's hotness to find hot topics. To reach this goal, the number of posts of each topic is taken as the topic's hotness. We choose the daily Top N topics with highest hotness value as the main topics of this study. Due to the presence of some advertising information in the extracted

hot topics, which are useless, it is experimentally concluded that when N=30 is chosen (N is not unique), a sufficient number of valid topics can be guaranteed. In this case, we extract the daily Top 30 hot topics for this study, and there are 9900 topics in total for eleven months. Because there is a possibility that a topic may be a hot topic for several days in a row, we are supposed to de-duplicate these 9900 topics and finally get 8359 unduplicated topics, which are used as our historical topic library (including curves and text contents) for this study. These topics are like “President Donald Trump announced that he and first lady Melania Trump has tested positive for COVID-19”, “Joe Biden just overtook Donald Trump in Pennsylvania, where he’s now leading by 5,594 votes”.

B. *Finding Topic-Related Posts*

In order to study the lifecycle curve of a topic, we need to know how hot the topic is on a daily basis. Thus, we need an algorithm to find the statistics of the number of posts a topic has on a daily basis, as the daily hotness of the topic.

By achieving this goal, we design a Topic-Finding-Posts algorithm. The algorithm performs keyword extraction from a library of posts to be searched to obtain keywords for each post, and then it calculates the Jaccard similarity coefficient score between the set of post keywords and the set of topic keywords to see if the post is similar to the topic or not. The higher the score, the more similar the two sets. Finally, it outputs the posts related to the topic.

Due to the sheer volume of computing and the limitations of machine, we are unable to find posts related to a topic for an entire year. Given that hot topics are generally not hotter than three months, we set the lifecycle to three months, the month in which the topic is located, the month before and the month after. For example, if a hot topic is on July 3, 2020, we would look for posts related to the topic in June, July and August of 2020, which means it requires us to calculate the Jaccard similarity coefficient score between the set of post keywords in these three months and the set of topic keywords to find topic-related posts. The flow chart of this algorithm is shown in Fig. 2.

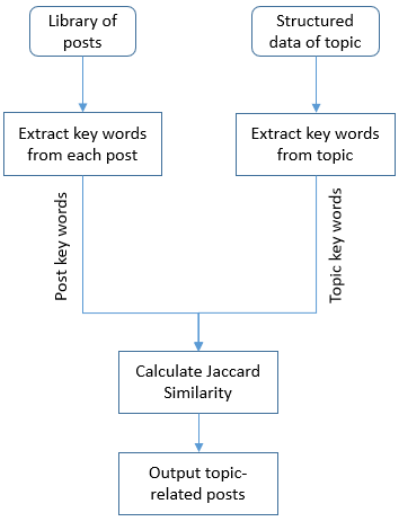


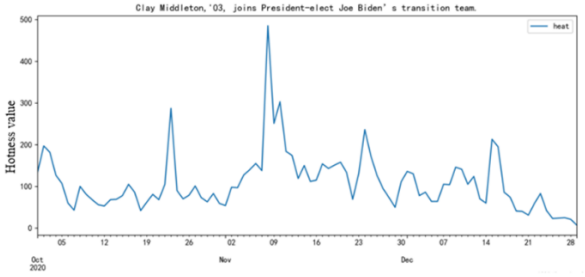
Figure 2. Flow chart of Topic-Finding-Posts algorithm.

C. *Drawing Historical Topic Lifecycle Graphs*

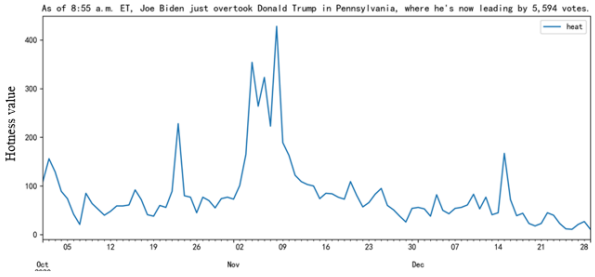
This step is to draw lifecycle graphs of all historical topics to build the historical topic library needed for this study. According to Section A in Chapter III, we know that there are a total of 8359 unique topics. The Topic-Finding-Posts algorithm of Section B in Chapter III is used to find posts related to these topics over a three-month period and the number is counted as the topic's hotness value. Using the topic's hotness value as the y-axis, the three-month span of time as the x-axis, and the topic description as the title, the lifecycle graphs of these topics are plotted and saved as a historical topic graph library, representing the lifecycle trends of hot topics that emerged during these eleven months.

D. *Topic Lifecycle Curve Shape Clustering Based on K-Shape*

Based on the results of Section C in Chapter III, we can obtain lifecycle graphs of thousands of historical topics. Since several hot topics appearing at the same time may be discussing the same thing, from this perspective they are actually one topic. For example, topic "Clay Middleton, ’03, joins President-elect Joe Biden’s transition team" and topic "Joe Biden just overtook Donald Trump in Pennsylvania, where he’s now leading by 5,594 votes" appeared in November 2020 are both about the US election, and they are similar in terms of their lifecycle curves. These two topics’ lifecycle curves are shown in Fig 3.



(a) Lifecycle curve of topic "Clay Middleton, ’03, joins President-elect Joe Biden’s transition team".



(b) Lifecycle curve of topic "Joe Biden just overtook Donald Trump in Pennsylvania, where he’ s now leading by 5,594 votes".

Figure 3. Examples of two topics that have similar lifecycle curves.

In response to this case, we decide to cluster topics with similar lifecycle curves into one class and form a curve that represents this class as the lifecycle curve for this class. There

are several advantages of doing this. First of all, it reduces the amount of computation since we only take the curve that represents each cluster into account. Secondly, it may help reduce the errors caused by the Topic-Finding-Posts algorithm on the hotness value of the topic lifecycle graph. Thirdly, it can eliminate some noise. For example, some oddly shaped graphs that appear only once are not representative, indicating that they are not common hot topics and will most likely not appear again in the future, which is not helpful for prediction, and these outlier topics can be found and discarded through clustering.

In this paper, K-Shape algorithm [12] is used for clustering, which is a clustering algorithm specifically for time series data and is concerned with similarity of shape. We use the tslearn package for clustering, which requires that the lengths of the different sequences should be the same. Thus, we cluster the curves by month, which ensures that the time series lengths of the lifecycle curves of topics in the same month are the same. In addition, we need to do feature scaling to bring all hotness values to the same magnitudes. To do this, we standardize the time series data by using z-normalization. Then, we use the normalized dataset to perform K-Shape clustering, where similar curves are clustered in one class and output centroid curves representing the lifecycle curves in this class. For example, the above lifecycle curves in Fig 3 are similar and can be clustered into one class, whose centroid curve is shown in Fig 4.

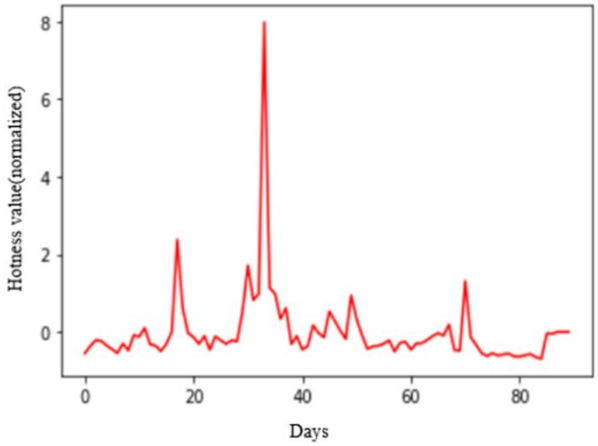


Figure 4. Example of centroid curve of a cluster (z-normalization).

All centroid curves after clustering are collected as a historical topic graph library. When a new topic emerges, the shape similarity between the new topic and the curves in historical topic graph library can be compared to predict the future trend of the new topic.

E. Calculation of Shape Similarity Based on DTW

DTW (Dynamic Time Warping) [13] is a dynamic programming algorithm that calculates the similarity of two time series [14], especially those of different lengths. When a new topic has been around for a while, we use this algorithm to calculate the shape similarity between the lifecycle curve of the new topic at that point and the centroid curves in the historical topic graph library in turn, and rank them to get some historical topic curves that are most similar to the current new topic. This gives an indication of some possible future trend directions for the new topic.

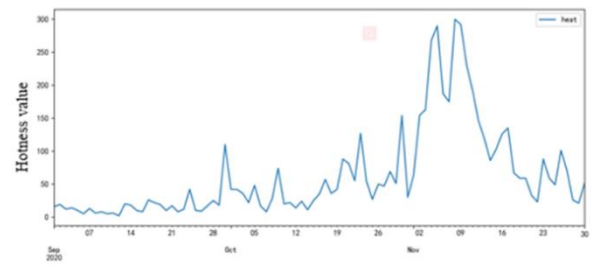
F. Calculation of Text Content Similarity Based on Keyword Matching

From Section D of Chapter III, we can see that we have successfully clustered topics with similar lifecycle curves and obtained the centroid curves representing each category of topics. Next, we perform keyword extraction for each category of topics to learn the main textual content. The Jaccard similarity coefficient score between these topic keywords and the new topic keywords are calculated in turn and ranked to find the curves of historical topics that are most similar to the new topic in terms of textual content. This gives an indication of some possible future trend directions for new topics when considering similarity in text content. When a new topic has just come out and there is no obvious curve, text content similarity can be considered to use to solve the cold start problem, but only for reference, as similar text content does not mean that the trend is similar.

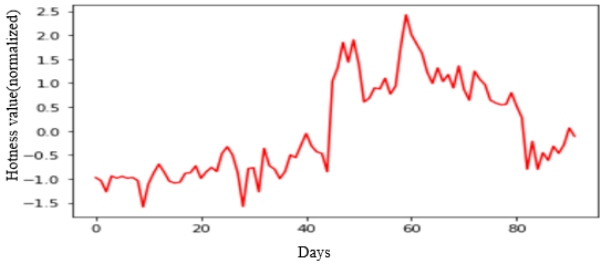
IV. Experiments and Evaluation

A. Topic Trend Prediction

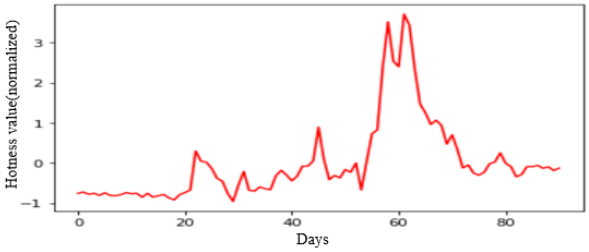
Following the order in Chapter III, we first extract the daily hot topics from the training set, then use the Topic-Finding-Posts algorithm to find the topic-related posts in a three-month cycle (here we set the similarity threshold of 0.45, if the Jaccard similarity coefficient score is greater than the threshold, the post is considered relevant to the topic). After that, we count the number of daily posts as the hotness to draw the lifecycle curve of the topic, and similar curves are clustered. Next, the DTW-based shape similarity calculation is performed on the clustered centroid curves and the test topic curves, and it is tested that when the similarity distance is less than 3.4, the trends of the two topics are similar. At the same time, the text content similarity calculation based on keyword matching is also performed (here the similarity threshold is also 0.45, and when the similarity is greater than 0.45, the text contents of the two topics are similar). Based on these experiments, historical topics that meet all the above conditions are considered similar to new topics, and their lifecycle trends can be used as a prediction of future trends of new topics. Let's take an example (see Fig 5), when the test topic "Trump's dream is America's dream, Biden's dream is China's dream, Ivanka says" shows a lifecycle trend (see Fig 5. a, the part of the diagram within the dotted line), we can get three similar curves from historical topic graph library by only considering shape similarity (see Fig 5. b c d). Then we consider text content similarity, only one curve meets all the conditions at the same time, which is the second one of the three predictions (see Fig 5. c). Then we get our predicted trend curve for the test topic.



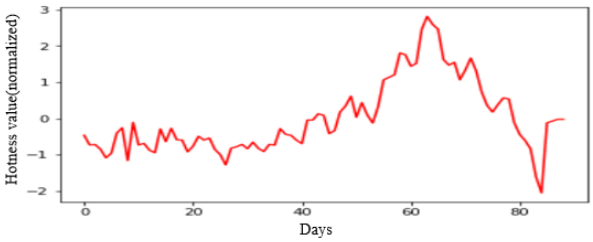
(a)



(b)



(c) the right one



(d)

Figure 5. Topic trend prediction experiments.

B. *Clustering Effect Evaluation – Silhouette Coefficient*

We use the Silhouette Coefficient as the effect evaluation index of this K-Shape clustering. The Silhouette Coefficient is a useful metric for evaluating clustering performance. It is computed by using mean distance between data points in the same cluster (cohesion) compared to the mean distance between data points in other clusters (separation) [15]. The calculated score ranges from -1.0 to 1.0. The higher the score, the better the clustering effect. To make the score computable, there have to be at least two clusters.

Assume that the data have been clustered into *k* classes. For data point $x(i) \in K$ (*K* is the cluster containing all the data points $x(i)$), $a_{x(i)}$ is the mean distance between $x(i)$ and every data point in the cluster *K*, $b_{x(i)}$ is the minimum mean distance between $x(i)$ and every data point in other clusters that is not a member of *K*. The calculation [16] of the Silhouette Coefficient of $x(i)$, the Silhouette Coefficient of each cluster, and the Silhouette Coefficient of all clusters can be shown as in (1), (2), and (3), respectively.

$$S_{(x_i)} = \frac{b_{x(i)} - a_{x(i)}}{max(a_{(x(i)}, b_{x(i)})} \quad (1)$$

where

$x(i)$ = data point in the cluster, $i = 1, 2, 3, \ldots, n$,

$a_{x(i)}$ = the mean distance between $x(i)$ and every data point in the cluster *K*, and

$b_{x(i)}$ = the minimum mean distance between $x(i)$ and every data point in other clusters that is not a member of *K*.

$$S_m = \frac{1}{n} \sum_{i=1}^{n} S_{(x_i)} \quad (2)$$

where

m = the number of the cluster, and

n = the number of data points in the same cluster.

$$S_{avg} = \frac{1}{k} \sum_{m=1}^{k} S_m \quad (3)$$

where

k = number of all clusters.

We take the data of November 2020 as an example and cluster out ten classes as shown below (see Fig 6), where the red line represents the centroid curve of each class with a S_{avg} of 0.5162703634739744. The results tell us that the clustering works well. Data from other months are treated in the same way.

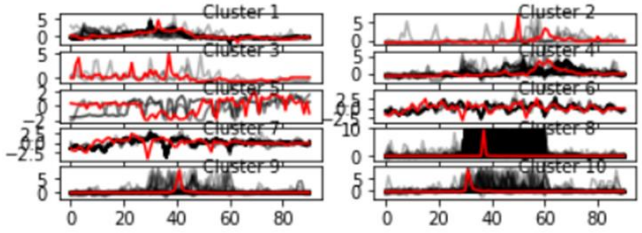


Figure 6. Clustering results of the data of Nov 2020(normalized).

C. *Prediction Effect Evaluation*

According to the clustering results of Section B in Chapter IV, we can briefly classify the type of clustering into short lifecycle class topics and long lifecycle class topics. The short ones refer to as above cluster 8, 9, 10, suddenly appeared to reach the peak and then the hotness immediately dropped and disappeared, mostly for some sudden events whose whole duration is just a few days. While the long ones are like cluster

5, 6, 7, whose hotness duration is long enough. They are usually serious events that need to be widely discussed. In this paper, we use the data of test set (120 topics in Jul 2021) as a collection of new topics. This method performs well on the test set, and the accuracy can reach 90%. For those test topics that are incorrectly predicted, we can see that there has not been a curve in the historical topic graph library similar to the curve of these test topics and there are no similar keywords in the text content either. In response to this situation, we add the lifecycle curves of these incorrectly predicted topics into the historical topic graph library to enrich it, so that these unmatched curves can be matched in the future.

V. Conclusion and Future Work

This paper proposes a topic lifecycle trend prediction algorithm based on Facebook data, which integrates shape similarity and text content similarity, and the results are more accurate than considering only shape similarity or only text content. The experimental results also show that the method is effective, but there are still some shortcomings that need to be improved.

1. For topics or lifecycle curves that have not appeared in history, it is impossible to make effective predictions, and the possible solution is to continuously expand the historical topic library to cover as many topics and curves as possible.
2. Because of the large amount of data, some topics have a very long lifecycle and the complete curve cannot be obtained. The algorithm can be optimized later to improve the running speed so that the complete lifecycle curve can be reached.
3. For shape similarity, the new topic needs to have a long enough lifecycle curve (to cover local or global features) to determine whether two topics are really similar by shape similarity calculation.

Acknowledgment

I would like to express my great appreciation and thanks to my supervisor Dr. Jun Shi, for her patient guidance and useful suggestions. I would also like to thank all my colleagues, for brainstorming with me and helping me solve the problems I encountered in this work. Finally, special thanks should be given to Shenzhen CyberAray Network Technology Co., Ltd, for supporting the data and resources I need.

References

[1] Kemp S. Digital 2020: July Global Statshot, Datareeportal.

[2] Jiang, Yue, Bojan Cukic, and Tim Menzies. "Fault prediction using early lifecycle data." The 18th IEEE International Symposium on Software Reliability (ISSRE'07). IEEE, 2007.

[3] Ishikawa, Shota, et al. "Hot topic detection in local areas using Twitter and Wikipedia." ARCS 2012. IEEE, 2012.

[4] Lu, Rong, and Qing Yang. "Trend analysis of news topics on twitter." International Journal of Machine Learning and Computing 2.3 (2012): 327.

[5] Zhao, Juanjuan, et al. "A short-term trend prediction model of topic over Sina Weibo dataset." Journal of Combinatorial Optimization 28 (2014): 613-625.

[6] Abuhay, Tesfamariam M., Yemisrach G. Nigatie, and Sergey V. Kovalchuk. "Towards predicting trend of scientific research topics using topic modeling." Procedia Computer Science 136 (2018): 304-310.

[7] Chen, Chengyao, et al. "Modeling scientific influence for research trending topic prediction." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. No. 1. 2018.

[8] Liu, Yunmei, et al. "The sustainable development of financial topic detection and trend prediction by data mining." Sustainability 13.14 (2021): 7585.

[9] Xu, Mingying, et al. "A scientific research topic trend prediction model based on multi-LSTM and graph convolutional network." International Journal of Intelligent Systems 37.9 (2022): 6331-6353.

[10] Ramos, Juan. "Using tf-idf to determine word relevance in document queries." Proceedings of the first instructional conference on machine learning. Vol. 242. No. 1. 2003.

[11] Mihalcea, Rada, and Paul Tarau. "Textrank: Bringing order into text." Proceedings of the 2004 conference on empirical methods in natural language processing. 2004.

[12] Paparrizos, John, and Luis Gravano. "k-shape: Efficient and accurate clustering of time series." Proceedings of the 2015 ACM SIGMOD international conference on management of data. 2015.

[13] Myers, Cory, Lawrence Rabiner, and Aaron Rosenberg. "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition." IEEE Transactions on Acoustics, Speech, and Signal Processing 28.6 (1980): 623-635.

[14] Meesrikamolkul, Warissara, Vit Niennattrakul, and Chotirat Ann Ratanamahatana. "Shape-based clustering for time series data." Advances in Knowledge Discovery and Data Mining: 16th Pacific-Asia Conference, PAKDD 2012, Kuala Lumpur, Malaysia, May 29-June 1, 2012, Proceedings, Part I 16. Springer Berlin Heidelberg, 2012.

[15] Kaoungku, Nuntawut, et al. "The silhouette width criterion for clustering and association mining to select image features." International journal of machine learning and computing 8.1 (2018): 69-73.

[16] Aranganayagi, S., and Kuttiyannan Thangavel. "Clustering categorical data using silhouette coefficient as a relocating measure." International conference on computational intelligence and multimedia applications (ICCIMA 2007). Vol. 2. IEEE, 2007.

Enriching RDF-based Document Management System with Semantic-based Reasoning

Maria Assunta Cappelli Ashley Caselli Giovanna Di Marzo Serugendo
CUI – Centre Universitaire d'Informatique
Université de Genève, Geneva, Switzerland
E-mail: maria.cappelli@unige.ch

Abstract

Paper-dependent companies spend most of their time organising and searching for documents, such as invoices, contracts, and budgets. To carry out those tasks, fiduciaries, insurance brokers, and other companies are pioneering Document Management Systems (DMSs). However, there is currently no DMS that allows the classification, understanding, and reasoning of a bundle of documents delivered by customers, and that helps companies create customer profiles to make better and faster decisions. Our proposal aims at easing the tasks of companies dealing with administrative documents of different kinds. We propose a semantic rule-based approach that permits to recognise and classify customers' documents, as well as to reason over those documents and create customer profiles based on the extracted information. We provide and discuss a case study grounded on the Swiss tax declaration. We developed a full Swiss tax ontology composed of 241 classes, as well as 120 semantic reasoning rules fully validated on the minimum set of administrative documents necessary to fill the Swiss tax declarations. Our approach automates activities related to the management of administrative documents, the profile of their clients, and the administrative documents they must deliver.

Index terms— Automatic Document Processing, Data Management Systems, Knowledge Graph-based Approach, Reasoning Engine

1. Introduction

Metadata-driven document management platforms have drastically simplified the work of document-dependent companies in the past decade. They enable professionals to find the right information, automate business processes, and enforce information control in any environment. Efficient management of documents is crucial not only for the optimal finding and use of documents but also for an effective and efficient work organisation. Gorelashvili et al. [5] note that in the legal sector, automated document management is essential to improve and streamline the way lawyers manage their practice. Document Management Systems (DMSs) ensure that documents are easily accessible, well-organised, and protected. Abbassova et al. [1] share the same opinion. They highlight the beneficial effects of DMSs on workflow forms, by automating the routing of documents between people, eliminating bottlenecks, and optimising business processes. They highlight the benefits of DMSs on the workflow assuming that the DMSs ensure more accurate organisation of business processes within the company through effective management and support the quality system in line with international norms, as well as efficient storage, management, and access to information and knowledge. However, the DMSs structure raises issues regarding the quality and completeness of the information sought, as the information is processed according to metadata. DMSs solutions either have considerable activation processes – and the associated fees – or are not well-suited for industry-specific professions. Also, since most solutions are business-based, there are currently no solutions for automating access to critical documents for individual consumers. There is not yet a DMS that allows the classification, understanding, and reasoning of customers' documents, automatically processing a bundle of customers' documents and creating a customer profile, in compliance to regulations. Tax declarations or insurance brokerage-related documents are still manually processed, transferred by e-mail or via consumers' cloud platforms.

This paper proposes a semantic rule-based approach for RDF-based DMSs. We designed a rule-based process that dynamically builds, reasons, and takes into account users' profiles and underlying regulations. This process is based on the information extracted from the documents users provide. Based on ontology capturing Swiss tax declaration, we designed rules on which the SHACL-based reasoner runs to derive inferences from the asserted RDF triples of

DOI reference number - 10.18293/DMSVIVA23-034

various tax households. A more detailed description of the approach can be found in the technical report [3].

The remainder of the paper is organised as follows. Section 2 provides an overview of existing approaches related to the DMSs. Section 3 describes the proposed approach and a case study. Section 4 shows our rule-based methodology. The SHACL-based implementation of the rules and their execution is shown in section 5. Finally, section 6 provides the evaluation of the defined rules, and Section 7 concludes the paper.

2. Related work

To handle the variety of documents written in different formats within an automated process, some works propose a semantic management approach for heterogeneous documents through the use of ontologies. They formalise the structure and interrelations of individual document types. These approaches monitor the process, take care of the various dependencies between documents, analyse the consequences of the changes made in one document on other documents, and engineer the synchronisation steps necessary to obtain a consistent document collection.

Motta et al. [8] propose an ontology-driven approach to enrich documents. This approach enables the development and integration of formal knowledge models with archives of documents. It extends what is currently available using "standard" information retrieval and search facilities by providing intelligent knowledge retrieval and additional knowledge-intensive services.

Fuertes et al. [4] developed an ontology for DMS concerning the construction field. The ontology aims at classifying documents along the life cycle of the research project, decreasing the interoperability and information exchange issues, establishing a hierarchical structure of the different areas that conform to the lifecycle of such projects, and finally enabling an interrelated system between these areas.

Doc2KG is a framework that provides a continuous conversion of open data to a knowledge graph, exploiting the existing domain ontology standards. The system handles the initial conversion of a DMS to a knowledge graph and supports the perpetual population of the created knowledge graph with new documents. The authors rely on a combination of NLP techniques to facilitate the information extraction and on constraint-solving techniques for knowledge graph creation and manipulation [12].

The Semantic Document Management System (SDMS) leverages a semantic approach for managing the lifecycle of semantic documents, from their authoring and publication to archival. The system allows defining documents as composite resources with document content units that are uniquely identified and semantically annotated. One relevant feature of the SDMS is the ability to share and exchange the document contents and semantics by the users [10].

Several other research deal with an SDMS and some of them cover relevant aspects of document modelling. However, none of them relies on semantic rules. For example, Yen-Hsien Lee et al. [7] develop a domain-specific ontology to support automatic document categorisation. The ontology includes a complete and detailed hierarchy of concepts that are used to represent documents related to information systems and technology as a set of concepts with relative weights. While scholars recognise the advantages of using an ontology with classes in terms of interpretability and understandability of classification decisions, no reference is made to the definition of semantic rules to make the use of the ontology more flexible. Sheng et al. [11] propose the use of ontology in the context of e-governance to model government data and create a semantic environment for managing government information. They present a semantic-based e-government system structure and use OWL as an ontology description language, which aims to provide the basis for data sharing and analysis. The authors detail the conceptual entity, conceptual property, and relationship between concepts that correspond respectively with class, property, and axioms in the OWL language. However, they do not model semantic rules to define constraints and restrictions on the data, ensuring that they are consistent with the ontological structure they have defined, or to exploit their reasoning power.

3. General approach

We propose a semantic rule-based approach for helping companies process (e.g. administrative) documents for their customers. Our proposal addresses the following research questions:

A) How to classify and multi-label a document based on extracted information providing its key features?

B) How to build and update clients' profiles based on the documents provided and the information extracted from those documents?

C) How can a reasoning process determine which documents the customers must deliver based on their profiles?

3.1. Case study

In this paper, we focus on the households' Swiss tax declaration, and the documents required to fill the tax declarations. We limit our case study to private households profiles composed of a single person, a widow/er, couples, households with or without children or any other dependent people, retired, or working. We also limited our

case study to the minimum set of administrative documents necessary to fill the Swiss tax declarations of the households described above, namely: yearly revenue, bank statements, health insurance policies and benefits, and family allowances, concerning every household member. We have included the health insurance statements as they are mandatory in Switzerland and anybody must provide them for tax declaration purposes.

We consider a scenario of a tax household consisting of two working parents with children. As each parent is employed, data is extracted from their two salary certificates. The data extraction process identifies the main features of the document that are necessary conditions for a document to be classified as a salary certificate. In response to the first research question, rules are applied to classify the documents as salary certificates based on the extracted features. The system then assigns a double tag of "*Tax*" and "*Income*" to the salary certificates. In response to the second research question, the system profiles the two parents as employees. Finally, in relation to the third research question, the system identifies other necessary documents that the two parents and their children must provide, such as health insurance.

3.2. Global workflow and architecture

Figure 1 depicts the global overview of the workflow of our approach, whose detailed description can be found in [3]. Such an architecture is composed of three modules: *(i)* actual documents (native PDFs or scanned documents) are processed through a *Document classification and information extraction module*. This module generates *JSON files for each document*, identifying its class (e.g., health insurance policy), as well as specific information extracted from the document (e.g., date, amount); *(ii)* assuming that the documents are identified as being part of the bundle of documents belonging to a specific tax household, the information extracted from the documents is also used to feed *JSON files profiles* of the household and its various members (e.g., widow/er, child, etc.); *(iii)* JSON information is then mapped to RDF by the *Reasoning, Labelling and Profiles updates* module, using an ontology for Swiss tax declaration, as well as people profiles. This module also contains a semantic rule-based reasoner, which serves on the one hand to update the profiles' information (e.g., health insurance policy for a new child means that child must be added to the household, possibly changing the household profile from couple without children to couple with children), and on the other hand to identify any missing document, based on the existing profiles, of the household (e.g., health policy or benefits are missing for a person identified as being a part of the household).

The details of the *Reasoning, Labelling, and Profiles update* module are the followings: *(i)* an *ontology of the Swiss tax* declaration terms and documents, based on actual official tax declaration legal documents and on actual documents needed to fill the tax declaration. The ontology defines concepts such as tax household, tax sections, documents, people, as well as profiles; *(ii)* the *rules*, defined for documents validation, updating the profiles based on new information, identifying missing documents (e.g. not provided in the bundle), and labelling documents; and *(iii)* the *RDF data* mapped from the JSON files (actual data) that contain information extracted from the documents using an information extraction process.

4. Rule-Based Methodology

We develop our semantically enriched DMS by designing semantic rules addressing the points A), B), and C) of Section 3.

Classification and multi-labelling rules. One or more label is assigned to each document to allow automated organisation of the documents into several predefined categories. Table 1 shows a *Salary Certificate* document as being multi-labelled as both a *Tax* and an *Income* document.

Customer profile rules. Infer the users' profile by analysing the documents they provided (*Document* → *User Profile*). We refer to them as **direct rules**. Table 1 shows that if a user delivers a *Salary Certificate*, then the user is tagged as being an *Employee*.

DOCUMENT ⟶ LABEL		
Salary Certificate	tagged as	*Tax*
Salary Certificate	tagged as	*Income*
DOCUMENT ⟶ USER PROFILE		
User	delivers *Salary Certificate*	*User* is an *Employee*
User	delivers *Health Insurance Policy*	*User* is an *Person*
USER PROFILE ⟶ DOCUMENT		
Employee	has to deliver	*Salary Certificate*
Person	has to deliver	*Health Insurance Policy*

Table 1: Examples of the defined rules

Documents delivery rules. Infer which documents match the profile of the clients (*User Profile* → *Document*). We qualify these rules as **inverse rules**. Table 1 shows that if a *Person* is tagged as *Employee*, then he/she must deliver a *Salary Certificate*. Additionally, any *Person* must deliver a *Health Insurance Policy* document.

5. Implementation

We implemented the aforementioned rules as inference rules using the SHACL language [6] – a W3C standard developed by the RDF Data Shapes Working Group for RDF graph validation[1]. It is a highly expressive language that lets

[1]RDF Data Shapes Working Group Charter, 2017, `https://www.w3.org/groups/wg/data-shapes/charters`

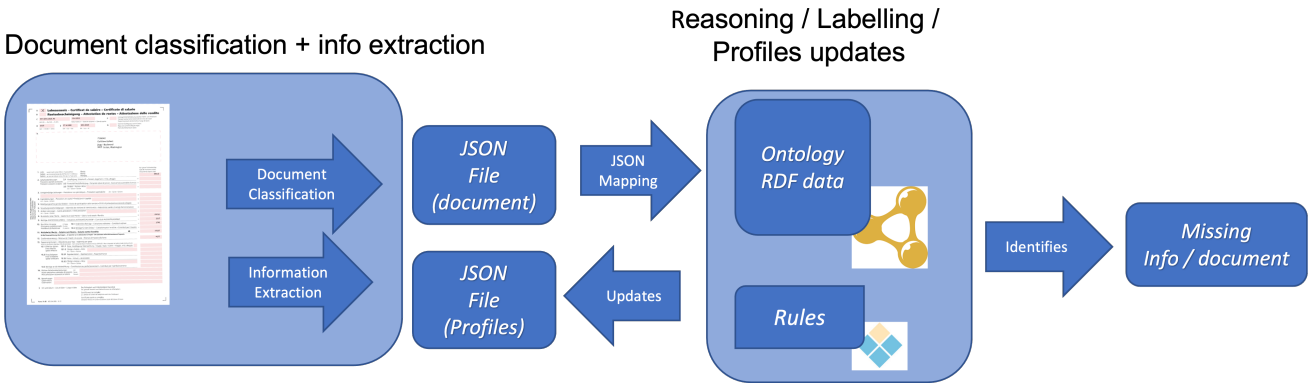


Figure 1: Overview of the global workflow as presented in [3]

its users write conditions, called shapes, that an RDF data graph must satisfy. In addition to the validation features, it also provides constructs for writing data transformations and inference rules through the SHACL Advanced Features vocabulary [2].

We created an ontology using Protégé [9] to represent the Swiss tax vocabulary. The ontology creation has been supported by extracting the domain's terms from the tax guide 2020 of the Geneva canton. Such an ontology defines a common vocabulary that will be used throughout the process to describe the *documents* and *user profiles*. As the system will be used in connection with tax declarations, tax items are also represented. Therefore, it defines classes for representing administrative and tax documents (e.g., salary certificates, bank account statements, health insurance, family allowances, etc...), and the users' profiles (e.g., married, single, in cohabitation, widow/er, divorced, separated, employee, self-employed, etc...). The ontology is composed of 241 classes, 24 data type properties, 615 axioms, and 15 object properties. Further information on the ontology can be found in [3].

After defining the ontology, we stated 92 property shapes and three sets of semantic rules (discussed in Section 3) resulting in a total of 120 rules, which included 78 multi-label rules, 21 customer profile rules, and 21 document delivery rules [3].

◇ FIRST TASK: A) *Classification of documents*

For each document a user may deliver, we identify its *sine qua non* elements. For instance, a *Salary Certificate* must include the following elements containing information about the person: employee's surname and first name, employee's address, employer, net and gross salary, etc... We then define a SHACL shape for each document type with the relevant elements the document must contain. Each element is represented as a SHACL property shape. For instance, as shown in Listing 1, a *Salary Certificate* document must contain: the employee's surname (`impots:PersonSurnameShape`) and first name (`impots:PersonFirstNameShape`), one and only one *Employer* (property named *"EmployerPropertyShape"*), and the amount (`impots:AmountShape`). By using such a SHACL shape, we can run a validation process that validates (or does not) the document as of the corresponding type. This can also be interpreted as "if the document contains all the *sine qua non* elements, namely the validation is positive, then it belongs to the specific class" and thus is assigned with that class.

```
impots:SalaryCertificateShape
  rdf:type sh:NodeShape ;
  sh:property [
      sh:path impots:employer ;
      sh:minCount 1 ;
      sh:maxCount 1 ;
      sh:class impots:Employer ;
      sh:name "EmployerPropertyShape" ;
    ] ;
  sh:property impots:AmountShape ;
  sh:property impots:PersonSurnameShape ;
  sh:property impots:PersonFirstNameShape ;
  sh:targetClass impots:SalaryCertificate ;
.
```

Listing 1: Relevant features of a *Salary Certificate* document represented as SHACL shapes

◇ SECOND TASK: A) *Multi-labelling documents*

Multi-labelling rules assign one (or more) label to each document. By using the assigned labels, the documents can then be automatically organised into predefined categories. We define such rules as SHACL inference rules. Their execution generates inferred triples of the form:

`< document impots:tag label >`

where *document* is the RDF individual of the document that is being labelled; impots:tag is a data property, defined in the ontology, for assigning the label to a document; and *label* is a string literal (xsd:string) containing the actual text value of the label. Listing 2 shows a rule labelling a document of type SalaryCertificate as both a tax document ("*Tax*" label) and an income document ("*Income*" label).

```
impots:SalaryCertificateShape
    rdf:type sh:NodeShape ;
    sh:rule [
        rdf:type sh:TripleRule ;
        sh:subject sh:this ;
        sh:predicate impots:tag ;
        sh:object "Tax" ;
    ] ;
    sh:rule [
        rdf:type sh:TripleRule ;
        sh:subject sh:this ;
        sh:predicate impots:tag ;
        sh:object "Income" ;
    ] ;
    sh:targetClass impots:SalaryCertificate ;
.
```

Listing 2: SHACL inference rule for labelling a document of type *SalaryCertificate*

In case there exist any triples in the data graph that fulfil the following conditions: *(i)* there exists a document (:documentX) in the data graph, and *(ii)* such document is of type *SalaryCertificate* (:documentX rdf:type impots:SalaryCertificate), therefore the execution of the rules shown in Listing 2 infers new triples shown in Listing 3.

```
:documentX impots:tag "Tax" .
:documentX impots:tag "Income" .
```

Listing 3: Triples inferred by the inference rule shown in Listing 2

◇ THIRD TASK: B) *Customer profile rules*

As the multi-labelling rules, the customers' profile rules are defined as SHACL inference rules. Listing 4 shows an example of such rules. Contrary to the example previously shown, where the targeted documents were all the RDF individuals of a defined class, this example shows an extended targeting condition expressed using the SPARQL language.

```
impots:SalaryCertificate_Employee-Shape
    rdf:type sh:NodeShape ;
    sh:rule [
        rdf:type sh:TripleRule ;
        sh:object impots:Employee ;
        sh:predicate rdf:type ;
        sh:subject sh:this ;
    ] ;
    sh:target [
        rdf:type sh:SPARQLTarget ;
        sh:prefixes impots: rdf: ;
        sh:select """
            SELECT ?this
            WHERE {
                ?sc rdf:type
                    impots:SalaryCertificate .
                ?sc impots:recipient ?this .
                ?this rdf:type impots:Person .
            }
        """ ;
    ] ;
.
```

Listing 4: SHACL direct rule inferring the *Employee* profile of a *Person* from the provided *SalaryCertificate*

The execution of the direct rule defined in Listing 4 infers new triples of the form:

```
< person rdf:type impots:Employee >
```

where *person* corresponds to the specific RDF individual; rdf:type is the property used to state that a resource is an instance of a class; and impots:Employee is the inferred class to which *person* belongs.

◇ FOURTH TASK: C) *Documents delivery rules*

Listings 5 and 6 show examples of user profile rules.

```
impots:EmployeeShape
  rdf:type sh:NodeShape ;
  sh:rule [
      rdf:type sh:TripleRule ;
      sh:subject sh:this ;
      sh:predicate impots:delivers ;
      sh:object impots:SalaryCertificate ;
    ] ;
  sh:targetClass impots:Employee ;
.
```

Listing 5: SHACL inverse rule inferring the need for an *Employee* profile to deliver a *SalaryCertificate*

The execution of the rule defined in Listing 5 infers new triples of the form:

```
< employee impots:delivers
impots:SalaryCertificate >
```

which means that any *Employee* must deliver a *SalaryCertificate*. The execution of the rule defined in Listing 6 infers new triples of the form:

```
impots:PersonShape
  rdf:type sh:NodeShape ;
  sh:rule [
      rdf:type sh:TripleRule ;
      sh:subject sh:this ;
      sh:predicate impots:delivers ;
      sh:object impots:HealthInsurance ;
    ] ;
  sh:targetClass impots:Person;
.
```

Listing 6: SHACL inverse rule for inferring that a *Person* profile needs to deliver a *Health Insurance*

```
< person impots:delivers
impots:HealthInsurance >
```

which in turn means that any *Person* must deliver a *HealthInsurance* policy document.

6. Evaluation

Concerning the evaluation of our approach we define two aspects:

- ◇ the performance of the information extraction process, which is addressed by precision, recall, and accuracy;
- ◇ the validation of the reasoning rules, which identifies missing documents or updates profiles in all cases.

Our use cases and tests are limited to a small set of documents, and the rules strongly depend on the quality of the information extraction process. Our aim with this work is to provide a proof of concept for the semantic-based approach, assuming the information is properly extracted from the documents. A more complete solution would need a larger dataset in order to be able to test all potential rules, as well as a thorough evaluation of the information extraction component.

Our focus is on the second aspect of the evaluation. Therefore, we validated all the rules regarding the points A), B) and C). Our technical report [3] provides a complete discussion of the rules validation. For testing purposes, we created synthetic data for various households and added them to a local RDF graph loaded into TopBraid Composer[2]. We used those data for testing the inference rules presented in Listing 2, 4, 5, and 6.

6.1. Evaluating multi-labeling rules

As we can see in Figure 2, the execution of the rules shown in Listing 2 infers two new triples that assign the two labels "*Income*" and "*Tax*" to the individual `impots:SalaryC12.3.334` which is of type *SalaryCertificate*.

[2] `https://www.topquadrant.com/`

impots:SalaryCertificate12.3.334	impots:tag	Income
impots:SalaryCertificate12.3.334	impots:tag	Tax

Figure 2: Inferred triples that assign two labels to a document of type *SalaryCertificate*

6.2. Evaluating users profile rules

We defined two individuals `impots:ZolaGiovanna` and `impots:Ladoumegue_Jules`. We assume that `impots:ZolaGiovanna` delivered a *SalaryCertificate* document. Based on the direct rule defined in Listing 4, since `impots:ZolaGiovanna` delivered such a certificate, the rule infers she is an *Employee*. Figure 3 shows the inferred triples.

impots:ZolaGiovanna	rdf:type	impots:Employee

Figure 3: The result of the execution of direct rules on an individual that delivered a document of type *SalaryCertificate*

Conversely, we defined `impots:Ladoumegue_Jules` as an *Employee*. Therefore, according to the inverse rule defined in Listing 5, the execution infers that since he is of class *Employee*, he must deliver a *SalaryCertificate* document. According to Listing 6, since he is also a *Person*, he needs to provide a *HealthInsurance* policy document. Figure 4 shows the mentioned inferences.

impots:Ladoumegue_Jules	impots:delivers	impots:SalaryCertificate
impots:Ladoumegue_Jules	impots:delivers	impots:HealthInsurance

Figure 4: The result of the execution of inverse rules on an *Person* with an *Employee* profile

7. Conclusion and future work

In this paper, we have presented a semantic rule-based approach for a semantically enriched DMS. The adoption of such an approach would ease performing tasks such as administrative document management, user profiling, and profiling-related ones. As depicted by Figure 1, the work presented in this article is a module that may be integrated into a wider solution.

In summary, we implemented a proof of concept based on the SHACL language as well as a functional prototype

that is rather complete on the reasoning, labelling, and profiling module. The implementation still needs to be validated on a large dataset of documents. We focused on the validation of the semantic-based reasoning rules.

Future work will take into account the dynamicity of the profiles (a person's profile might change over time) as well as the integration of such a module into a wider DMS service.

Acknowledgments

This research was supported by Innosuisse within the framework of the innovation project 50606.1 IP-ICT "Addmin". The authors thank Anne-Françoise Cutting-Decelle, Assane Wade, Claudine Métral, Gilles Falquet, Graham Cutting, and Sami Ghadfi for their valuable collaboration with the "Addmin" project.

References

[1] V. Abbasova. Main concepts of the document management system required for its implementation in enterprises. *ScienceRise*, 1:32–37, 02 2020.

[2] D. Allemang, S. Steyskal, and H. Knublauch. SHACL advanced features. W3C note, W3C, June 2017.

[3] G. Di Marzo Serugendo, G. Falquet, C. Metral, M. A. Cappelli, A. Wade, S. Ghadfi, A.-F. Cutting-Decelle, A. Caselli, and G. Cutting. Addmin: Private computing for consumers' online documents access: Scientific technical report. 2022.

[4] A. Fuertes, N. Forcada, M. Casals, M. Gangolells, and X. Roca. Development of an ontology for the document management systems for construction. In *Complex Systems Concurrent Engineering*, pages 529–536. Springer, 2007.

[5] L. Gorelashvili. The importance of digitalization of legal documents preparing process and its impact on peoples' legal guarantees. In R. Geibel and S. Machavariani, editors, *Digital Management in Covid-19 Pandemic and Post-Pandemic Times*. Springer, Cham, 2023.

[6] H. Knublauch and D. Kontokostas. Shapes constraint language (SHACL). W3C recommendation, W3C, July 2017.

[7] Y.-H. Lee, P. J.-H. Hu, W.-J. Tsao, and L. Li. Use of a domain-specific ontology to support automated document categorization at the concept level: Method development and evaluation. *Expert Systems with Applications*, 174:114681, 2021.

[8] E. Motta, S. B. Shum, and J. Domingue. Ontology-driven document enrichment: principles, tools and applications. *Int. J. Hum. Comput. Stud.*, 52(6):1071–1109, 2000.

[9] M. A. Musen. The protégé project: a look back and a look forward. *AI Matters*, 1(4):4–12, 2015.

[10] S. Nesic, D. Gasevic, and M. Jazayeri. Semantic document management for collaborative learning object authoring. In *2008 Eighth IEEE International Conference on Advanced Learning Technologies*, pages 751–755. IEEE, 2008.

[11] L. Sheng and L. Lingling. Application of ontology in e-government. In *2011 Fifth International Conference on Management of e-Commerce and e-Government*, pages 93–96. IEEE, 2011.

[12] N. Stylianou, D. Vlachava, I. Konstantinidis, N. Bassiliades, and V. Peristeras. Doc2kg: Transforming document repositories to knowledge graphs. *Int. J. Semantic Web Inf. Syst.*, 18(1):1–20, 2022.

DSWA: A Dilated Shift Window Attention Method for Chinese Named Entity Recognition

Xinyu Hou, Cui Zhu, Wenjun Zhu

School of Computer Science, Faculty of Information Technology, Beijing University of Technology, Beijing, China

E-mail: houxy18@emails.bjut.edu.cn, {cuizhu, zhuwenjun}@bjut.edu.cn

Abstract

In recent times, numerous models tried to enhance the performance of Transformer on Chinese NER tasks. The model can be enhanced in two ways: one is combining it with lexicon augmentation techniques, the other is optimizing the Transformer model itself. According to research, fully connected self-attention may scatter the attention distribution, which is the reason for worse performance of the original Transformer with self-attention. In this paper, we attempt to optimize the Transformer model especially attention layer. Therefore, a novel attention mechanism, Dilated Shift Window Attention, is proposed to address this problem. By using Window Attention, this method improves the model's capacity to deal local information, meanwhile, the model can still manage long text and long-distance dependencies owing to the Window Dilatation mechanism. Experiments on various datasets also show that DSWA replacing fully connected self-attention improves the model's performance on the Chinese NER task.

Keywords— NLP, Chinese NER, Transformer, Window Attention, Self-attention

1. Introduction

Named entity recognition (NER)[1], an upstream task of natural language processing (NLP)[2], tries to extract entities from the original input text, such as names of people, places, and organizations, etc. It plays an integral role in many downstream NLP tasks.

Transformer[3] has recently excelled at other NLP tasks such as text translation, but in the domain of Chinese NER, the Transformer with self-attention produces unsatisfactory results. As a result, it is more frequently used in combination with other approaches or models. Comparatively speaking, a number of models based on CNN[4] or LSTM[5] have produced favorable results, especially BiLSTM+CRF[6] has been a proven NER solution in daily applications since its stability.

The majority of NLP tasks require researchers to broaden the model's receptive field in order to understand the semantics of sentences or paragraphs. This is also the reason why models are excel in addressing long text comprehension or long distance dependency. A good example is the self-attention mechanism, which, by calculating the attention of the entire vector sequence, enables the model to effectively perceive a wider range of data. But nevertheless, entities, the recognition targets for NER task, frequently appear in the text as a short neighboring sequence like a word or phrase. It is crucial to put more attention to local areas of various ranges and understand the semantics of sentences as well. For example in Figure 1, two entities occupy only a small part of the sentence, attention in local areas may handling them well. Meanwhile, both of them contain the same word "Zhongshan", the first one is a person name and the second one is a place, in order to discrimination, model need the ability to understand the semantics of sentences. As a result, the self-attention that acting on the entire vector sequence may good at understanding, but weak in processing such local information, which may be the reason why the Transformer model with self-attention does dissatisfiedly on Chinese NER task.

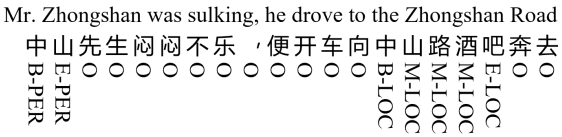


Figure 1: An Example of Chinese NER

In this paper, Dilated Shift Window Attention (DSWA), a novel attention approach, is proposed to solve the following issues. We investigated the reason why Transformer models with self-attention perform poorly on Chinese NER tasks. Additionally, we argue that the superior local information

10.18293/DMSVIVA2023-106

perception capacity of LSTM and CNN, compared to Transformer, make them more suitable for Chinese NER tasks. The new attention approach aims to improve the model's perception of local information while preserving as much of its capacity for long texts as possible. The effectiveness of this novel attention mechanism in enhancing the model's performance on the Chinese NER problem was empirically confirmed.

2. Related Work

2.1. LSTM

The Long-Short Term Memory Network (LSTM), which excels at handling serialized input and has had success in NER tasks. LSTM uses a gate mechanism to parse text as a sequence. Each cell in model processes a character while gathering data from the prior context.

Zhang proposes the lattice-LSTM[7] for the Chinese NER task, which enhances the adaptability of the LSTM model for the Chinese NER task through including lexicon information into the character-level LSTM model and adding an additional cell to encode lexicon information. However, the way the model introduces lexicon information reduces its parallelization efficiency. Hence, Liu proposed the WC-LSTM[8] in an effort to make improvements, so that the lexicon representation brought in for each character is static and stationary, improving the model's parallelization capability.

Furthermore, compared to self-attention, this method is able to process nearby contextual data more effectively. As a result, the LSTM and Transformer in conjunction can supplement the self-attention's ability to comprehend local information[9].

2.2. CNN

In the Chinese NER task, convolutional neural network (CNN) and its modifications also perform well. When dealing with brief text sequences like entities, CNN has a distinct advantage due to its high local information sensing. When applied to the Chinese NER task, CNN is enhanced by the external lexical information, meanwhile adjusting the weights of various words using the feedback layer; this is the LR-CNN[10] proposed by Gui. As a result, this could satisfactorily handle the issue of conflicting lexical information caused by sentence breaks.

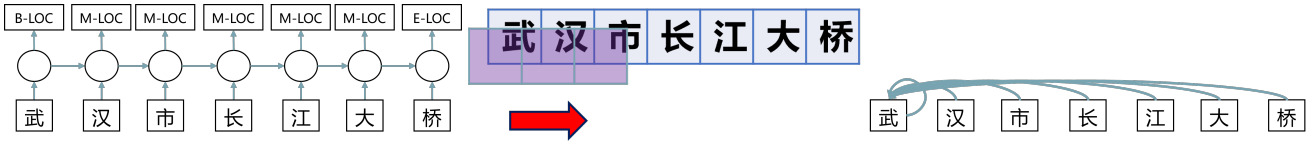
On the other hand, an effort is made to maintain the benefit of CNN's local information sensing capability while also enhancing its global information sensing capability in order to better address the issues associated with lengthy entities and long-distance dependency. With Iterative Dilated Convolution (IDCNN)[4], adding holes to the convolution kernel widens the model's receptive field while maintaining a constant computation. Additionally, IDCNN considers the iterative component. As the number of iteration rounds increases, the holes in the dilated convolutional kernel gradually enlarge and the perceived field exponentially expands. IDCNN requires less computing than traditional pooled convolution and is better at handling global data. It also gave us a fresh concept for enhancing the self-attentive mechanism.

2.3. Transformer

The traditional Transformer[3] model with self-attention performs poorly on the Chinese NER task, in contrast to other NLP domains, despite numerous improvements, which nevertheless lead to a decent score. An excellent illustration is the Simple-Lexicon model, which provides the Transformer with external lexical data. In order to adapt word embeddings to the task features of the Chinese NER, researchers have adopted a technique called "Soft-lexicon"[11] to do so. This method has produced promising results. A lexicon itself is a series of characters used in a variety of contexts, therefore the information contained inside it also serves as local information. The addition of lexicon data gives the model a new way to look at data processing and fills in the gaps left by the character-level self-attention method.

Since the majority of Chinese NER employ character-level models, this approach to applying additional lexicon, known as lexicon enhancement, has several applications and is used by many of the models mentioned above. A popular area of study in recent years has been the use of Transformer and lexicon enhancement in combination. Generally speaking, there are two technical approaches to lexicon enhancement. The first is termed dynamic architecture, and it seeks to create a dynamic architecture that is suitable with lexicon information and character embedding. The FLAT[12] model, which allows Lexicon data to be input into the model along with the character embedding, is a typical example. The Soft-lexicon discussed above is an application of the second method, which is called Adaptive Embedding, it integrates lexicon data into the character embedding without changing the input. Furthermore, because the DSWA replaces self-attention without altering the input, it works well with both two lexicon augmentation approaches. Both the FLAT model employing Dynamic Architecture and the Transformer model combined with Soft-lexicon have been enhanced after applying our DSWA, as you will see in the experimental part that follows.

Besides, other studies have attempted to enhance fully connected self-attention, Biao Hu et al. presented Adaptive Threshold Selective Self-Attention(ATSSA)[13] as such an approach. This method establishes a threshold for the at-



(a) LSTM (b) CNN (c) Self-Attention

Figure 2: Schematic for LSTM, Convolution, and Self-attention

tention score, and only the vectors that are higher than this threshold are used in the computation of attention, while the rest being discarded. In contrast to fully connected self-attention, this strategy constricts the attention distribution, but it's hard to set an appropriate threshold, may omit some information. Besides, the figure of attention visualization in this thesis demonstrates that, attention tends to be focused on the adjacent context in Chinese NER. It provides theoretical support for our approach that based on window attention.

In addition, the Bert-LSTM model is a proven industrial solution for Chinese NER tasks. The Transformer-improved Bert model has become a popular pre-training model for Chinese NER tasks. In addition to enhancing the Transformer model itself, it is also commonly used in combination with other models; a prime example of this is the LSTM-in-Trans model mentioned above. This can significantly increase the model's capacity for sequence modeling.

3. Method

As previously mentioned, the LSTM and CNN is adept in handling a limited range of input. Contrarily, the Transformer model is useful for dealing with long-range dependencies due to the self-attention mechanism used by the original Transformer, which computes attention for each vector with all vectors globally. For instance, Figure 2 is a schematic diagram of comparison between the three models. However, the entities to be identified by the NER task are frequently concentrated within the range of a few words, full-connected self-attention may not good at dealing in such a small range. Therefore, it is a challenge that how to limit the range of attention in order to deal local information on Chinese NER task.

Taking into account the mentioned considerations, we attempt to explore a strategy for Chinese NER task that better dealing with local information. It is worth noting that the window attention approach has shown effective results in other deep learning domains, effectively extracting local features from the input data. For instance, when performing object detection[14] tasks, the objects to be detected do not take up the whole image, the usage of the Window Attention method enables the neural network to concentrate on a specific area of the image to help extract reliable information. This led to the conclusion that window attention would be a useful tool for obtaining information about entities, since the process of identifying things in sentences for the NER task is quite similar to this.

However, there is a clear issue if the model using Window Attention alone: how can windows communicate with one another? If the window size is two, all words are divided into distinct windows, as in Figure 3 for example. Chinese words are made up of individual characters, thus it is absurd if characters that make up a word were split into different windows, and unable to communicate with others.

Figure 3: Words separated into different windows

The Shift Window Attention served the solution for this problem, which involves shifting the window after doing a round of Window Attention and then perform a subsequent round of attention computation in new window. Following the completion of the attention calculation inside each window, the attention calculation window is shifted by a predetermined amount, which does not exceed the window's size, and a new attention calculation is then carried out inside the relocated window. Similar to Figure 4, the attention calculation window is shown in red and moves between two rounds of attention computation. In this manner, the vectors situated in different windows can interact with others.

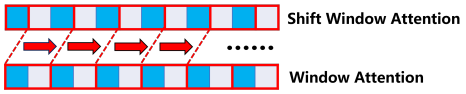


Figure 4: Shift Window Attention

In addition, if a long entity appears in the text that exceeds the length of a window, also a few of long-distance dependencies need to be solved, it cannot be handled by

simply using Window Attention. For these reason, although an approach that focus attention computing more locally is expected, the capability to obtain information on a larger scale is still necessary. Dilated Convolution may be an good idea for solution while considering about this issue, since it uses a "holey" convolution kernel to increase the receptive field of the convolution kernel. Similarly, it may allow for the expansion of the receptive field of the window when computing attention.

Each vector that occupies the same location in each attention calculation window is take out to create a number of new vector sequences, then another attention calculation is performed with the new sequences. The first vector within each window generates a new sequence, as shown in the Figure 5, and the other vectors does the same. If the window size is two, two new vector sequences are then processed similarly and iterated constantly to achieve cross-unit interaction.

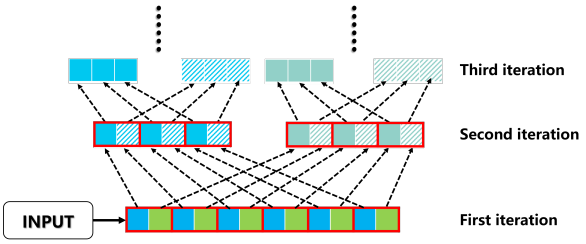


Figure 5: Window Dilitation

Due to the window attention and shifted window attention, each vector in the new sequence contains information about its nearby vectors. Thus, each vector that makes up the new sequence can be regarded as a representative of numerous vectors in the original sequence, the receptive field of one attention window in the new sequence is equivalent to several windows in the original sequence. As shown in Figure 6, several iterations are carried out in this manner, as the number of iterations rises, the receptive field of the attention window exponentially grows. This approach is named as Window Dilatation because a new window seems to be composed of original sequence with holes, similar to the Dilation Convolution.

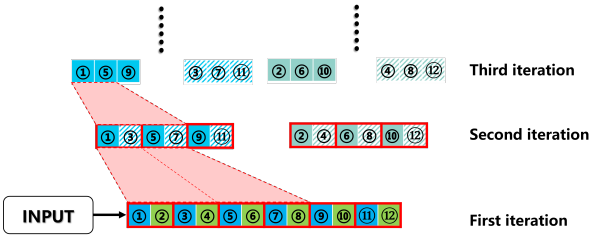


Figure 6: Receptive field changing in different iterations

The frame of our new attention module is a multi-round iteration and separated into three parts in each iteration: Window Attention, Shifted Window Attention, and window dilatation. The model can concentrate on local information by using Window Attention and Shifted Window Attention. Additionally, it enables each vector to carry the data of other vectors that come either before or after it. In this manner, the new sequence retains as much of the old sequence's information as feasible during the window dilatation. The iteration is then continued concurrently through the various sequences acquired after the window dilatation. Once a certain number of rounds have been completed or the new sequence is no longer lengthy enough to enable window dilatation, the iteration is stopped. What's more, neighboring vectors are split off into various sequences by window dilatation in the initial iterations, doing an additional Shift Window Attention at the end of the iteration, enable interaction between the original nearby vectors.

The most unique feature of DSWA is that concentrating more on extracting the local information than the self-attentive mechanism does, meanwhile information in a lager range can be fused partly. Our observation is that entities frequently exist in sentences as a string of words, therefore our NER task will be made easier by using this method to gather local information. In actuality, after experimental validation, our new attention module does boost Chinese NER's accuracy, particularly on the Weibo dataset where other approaches struggled.

4. Experiment

Weibo[15], MSRA[16], and Resume[7], three open datasets frequently utilized for Chinese NER tasks, are employed in this investigation. The Weibo dataset was gathered from the Chinese social networking site Sina Weibo; the MSRA, created by Microsoft from the news domain; and the Resume, gathered from the resumes of Chinese businesspeople. Statistics of the above datasets are shown in Table 1. Our method significantly improves the result of Weibo dataset.

Table 1: Statistics of datasets

Datasets	Type	Train	Dev	Test
Weibo	Sentence	1.4k	0.27k	0.27k
	Char	73.8k	14.5k	14.8k
Resume	Sentence	3.8k	0.46k	0.48k
	Char	124.1k	13.9k	15.1k
MSRA	Sentence	46.4k	-	4.4k
	Char	2169.9k	-	172.6k

Figure 7 is the frame of our model. In summary, it's a typical Transformer model, whereas the self-attention layer

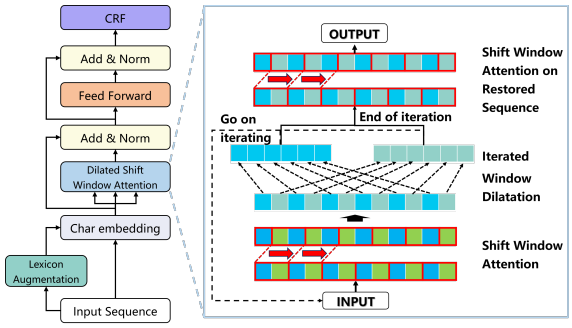


Figure 7: The Frame of DSWA

is replaced by our DSWA layer. It is worth noting that no matter which way of lexicon augmentation is used, there is a superior improvement after utilizing DSWA. Two representative models of both approaches are chosen in the experiment, FLAT is a typical example of dynamic architecture, while SoftLexicon is adaptive embedding.

SoftLexicon(Transformer) and FLAT model both is an enhanced version of the Transformer, by including external Lexicon information, the two model are enhanced considerably. Since self-attention mechanism were used by both of them, we can replace the self-attention layer with our DSWA to verify the effectiveness of our method.

Table 2: Performance on Three Datasets (F1 Score)

Model	Weibo	Resume	MSRA
Lattice-LSTM	58.79	94.46	93.18
TENER[17]	58.17	95.00	92.74
BERT+LSTM+CRF	67.33	95.51	94.83
FLAT	68.55	95.86	96.09
SoftLexicon	70.50	95.54	95.87
SoftLexicon + DSWA	**72.12**	**96.72**	**96.25**

The Table 2 shows that the DSWA-based model produced better results, particularly for the Weibo dataset where the improvement was greatest. The Weibo dataset, which is taken from the social media platform, has a strong propensity to be colloquial, meanwhile, there are a big number of short sentences and a higher occurrence of short entities according to our analysis. Therefore, DSWA, which is good at handling local information, is more suitable to handle such dataset. Both the MSRA and the Resume are plainly more stringent in their language, and more longer entities in them. Therefore, Weibo has been most significantly improved by our DSWA strategy, which places a strong emphasis on local information.

The models selected in the table are all classical models for Chinese NER task. Lattice-LSTM was the first to propose the method of external lexicon, Tener model was the first to improve Transformer, making model achieved good results on NER tasks, LSTM-CRF is a well-known baseline model for NER task, and both Flat and SoftLexicon (Transformer) are classical models that combine Lexicon Augmentation with Transformer as described above.

Table 3: Comparison of Self-Attention and DSWA (F1 Score)

Model	Weibo	Resume	MSRA
FLAT	68.55	95.86	96.09
FLAT + DSWA	69.63	96.21	96.03
SoftLexicon	70.50	95.54	95.87
SoftLexicon + DSWA	**72.12**	**96.72**	**96.25**

As shown in Table 3, comparing two Transformer-based models, the DSWA version have improved over the self-attention version; once more, the improvement is particularly noticeable on the Weibo dataset. It has been proven that using the DSWA method increases the model's accuracy and offers it a greater local information gathering capabilities. Therefore, we believe that DSWA can be used to replace the self-attention layer on more self-attention-based Transformer models to improve the model on Chinese NER task.

Concerning the problem of computation size, even though our method computes attention in multiple rounds, the computation size of the model actually decreases. When one attention computation is viewed as a unit, each vector of self-attention performs an attention calculation with every other vector. A total of n^2 attention computations are needed, where n is the length of the sequence produced by the sentence translation into vectors. However, Window Attention only needs to calculate attention within the window, and each window needs to perform W^2 attention calculations, where W is the length of the window, there are n/W windows in the sentence, so a sentence only needs to perform $W^2 \times n/W$, i.e. $n \times W$ attention calculations. Cause the window length W must be shorter than the sentence length n, the computation's complexity stays the same or even less than for self-attention. Even after r rounds of iteration, $r \times n \times W$ is smaller than n^2. In practice, r is generally set to a maximum of 4, because in the 4th iteration, the window's receptive field expands to 16, a size that essentially exceeds the length of all entity.

Table 6 demonstrates that the DSWA approach is marginally more efficient than doing a single round of self-attention computation, even numerous rounds of attention computation were carried out.

Compared with ASSTA, another method that modifies attention mechanism, in comparison, DSWA gets a similar f1 score, but a significant improvement in computational complexity, as shown in Table 4 and Table 5.

Table 4: Performance of ATSSA and DSWA

model	Weibo	Resume	MSRA
FLAT+ATSSA	72.53	96.73	96.45
SoftLexicon + DSWA	72.12	96.72	96.25

Table 5: Complexity of different methods. r is the rounds of iteration, W is the length of the window, r, W are constants.

Model	Complexity
Self-attention	O(n^2)
ATSSA	O(n^2 + 2n + nlogn)
DSWA	O($r \times W \times n$)

5. Conclusion

Dilatation Shifted Window Attention, which we suggest as a new attention calculation approach in this paper, is intended to improve the Transformer model's perception of local information while still preserving some perception of global information. It can better receive local information thanks to the use of the Window Attention mechanism, yet it can still perceive global information according to the iterative window dilatation approach, which enlarges each window's receptive field.

In brief, DSWA is a way of attention calculation that may be employed in place of self-attention and has a better ability to gather local information than the self-attention approach. It may also do well on other tasks that need deal data more locally. The experiment's findings match what we had anticipated.

Acknowledgement

The work is supported by the National Natural Science Foundation of China under grant 62276011.

Table 6: Efficiency of Two Attention Method

Datasets	Average Seconds per epoch	
	DSWA	Self-attention
Weibo	44	64
MSRA	2020	2729
Resume	105	131

References

[1] Jason P.C. Chiu and Eric Nichols. Named Entity Recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370, 07 2016.

[2] Detmar Meurers. Natural language processing and language learning. *Encyclopedia of applied linguistics*, pages 4193–4205, 2012.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[4] Emma Strubell, Patrick Verga, David Belanger, and Andrew McCallum. Fast and accurate entity recognition with iterated dilated convolutions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2670–2680, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[5] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[6] Ying An, Xianyun Xia, Xianlai Chen, Fang-Xiang Wu, and Jianxin Wang. Chinese clinical named entity recognition via multi-head self-attention based bilstm-crf. *Artificial Intelligence in Medicine*, 127:102282, 2022.

[7] Yue Zhang and Jie Yang. Chinese ner using lattice lstm. *arXiv preprint arXiv:1805.02023*, 2018.

[8] Wei Liu, Tongge Xu, Qinghua Xu, Jiayu Song, and Yueran Zu. An encoding strategy based word-character lstm for chinese ner. In *North American Chapter of the Association for Computational Linguistics*, 2019.

[9] Jun Yin and Cui Zhu. Embedding lexicon and direction information in chinese ner. In *International Conference on Computer Information Science and Artificial Intelligence*, 2021.

[10] Tao Gui, Ruotian Ma, Qi Zhang, Lujun Zhao, Yu-Gang Jiang, and Xuanjing Huang. Cnn-based chinese ner with lexicon rethinking. In *International Joint Conference on Artificial Intelligence*, 2019.

[11] Minlong Peng, Ruotian Ma, Qi Zhang, and Xuanjing Huang. Simplify the usage of lexicon in chinese ner. *ArXiv*, abs/1908.05969, 2019.

[12] Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. FLAT: Chinese NER using flat-lattice transformer. In *Proceedings of the 58th Annual Meeting of*

the Association for Computational Linguistics, pages 6836–6842, Online, July 2020. Association for Computational Linguistics.

[13] Biao Hu, Zhen Huang, Minghao Hu, Ziwen Zhang, and Yong Dou. Adaptive threshold selective self-attention for Chinese NER. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1823–1833, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[14] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, pages 1–20, 2023.

[15] Nanyun Peng and Mark Dredze. Named entity recognition for Chinese social media with jointly trained embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 548–554, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[16] Gina-Anne Levow. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 108–117, Sydney, Australia, July 2006. Association for Computational Linguistics.

[17] Hang Yan, Bocao Deng, Xiaonan Li, and Xipeng Qiu. Tener: Adapting transformer encoder for named entity recognition. *ArXiv*, abs/1911.04474, 2019.

Consistency analysis of UML models*

Guo-Fu Tang[1,2], Jian-Min Jiang[1,2], Hao Wen[3,4]
[1]Automatic Software Generation and Intelligence Service Key Laboratory of Sichuan Province
[2]Chengdu University of Information Technology, Chengdu 610103, China
[3]Chengdu Institute of Computer Applications, Chinese Academy of Sciences, Chengdu 610103, China
[4]University of Chinese Academy of Sciences, Beijing 101408, China
domcofu@qq.com; jjm@cuit.edu.cn; wenhao21@mails.ucas.ac.cn

Abstract

The inconsistencies of UML models (diagrams) during the software development process may cause errors in documents or programs. Massive consistency rules have been proposed by researchers to detect inconsistencies in existing works. However, the approaches of consistency analysis have not been addressed adequately in literature. In this paper, we propose a new approach for analyzing consistencies between UML models. We formally specify UML models and define the consistency relation between UML models. Based on the consistency definition, we first discuss the composition and decomposition of consistencies, and then explore the equivalence of consistencies.

Index Terms UML model, Consistency, Composition, Equivalence

1 Introduction

To develop a modern software system in the Model-Based Engineering approach, multiple perspectives of modeling the system are necessary. The UML is a semi-formal graphical modeling language [15], and is widely used in Model Driven Engineering (MDE) due to its multiple perspectives for describing software systems. Software artifacts (e.g., software architecture and implementation codes) are interrelated to UML diagrams in whatever versions, levels of abstraction, and stages. Consequently, it is often unable to avoid faults because of differences between UML diagrams in software development. Specifically, the evolution of the models or diagrams is frequently accompanied by augmenting, reducing, or modifying, which potentially results in contradictory specifications (inconsistencies). In order to classify such conflicts, there are seven UML consistency dimensions in the systematic researches [23, 8, 22, 11, 2], and a binary relation can capture these dimensions, including those that refer to *endogenous consistency* [14].

Multiple UML diagrams need to be strictly compliant in order for a software system to be accurately and completely described, especially for Safety Critical Systems (SCS). Without complete and consistent design models, programmers need to manually supplement the lack of design models with codes, which may cause faults and insecurities. Thus, we proposed to design accurate models because the models are easier to be analysed in formal methods. Corresponding to categories of UML diagrams, consistencies between these diagrams are generally divided into structural(i.e., syntactic) and behavioral(i.e., semantic) ones that has been systematically investigated in research [18, 19]. The consistency rules are sophisticated and cannot distinguish between binary and N-ary relations (i.e., the number of UML models involved in consistency rules is uncertain). For example, rule 15 from [18] involves a class diagram, a state machine diagram, and a activity diagram. Approaches to check N-ary consistencies have not yet been fully developed and the relevant theories are merely reviewed as well [18]. While using multiple rules to detect inconsistencies in a large system, it is obviously arduous to manage the duplication or absence of rules. Thus, it is necessary to present binary consistency relations for unifying N-ary consistency relations between UML models.

There are many existing efforts contributing to managing consistency relations or rules among UML models whereas ignoring the relationships between the rules [1, 17, 13]. In this paper, we propose a novel approach for analyzing consistencies between UML models. We formally spec-

*This work is supported by National Key R&D Program of China (No. 2022YFB3305104), National Natural Science Foundation of China (No. 61772004), and Scientific Research Foundation for Advanced Talents of Chengdu University of Information Technology (No. KYTZ202009).
DOI:10.18293/DMSVIVA2023-187

ify UML models and define the consistency relation between UML models. Based on the consistency definition, we first discuss the composition and decomposition of consistencies, and then explore the equivalence of consistencies. Because of such a duality of consistency relations, our method saves complexity and makes consistency characteristics more extensible, and formal methods aid in removing ambiguities and enforcing consistency. Our work aims at describing and managing consistency in complete UML diagrams for a system. Due to the space limitation, all proofs are deleted from our paper.

2 Model Composition

In this section, we will introduce a formal model [20] to specify UML diagrams, and then show characteristics of the formal model. In UML [15] various software artifacts are all regarded as models and its constituent parts are model elements. Such consideration facilitates the analysis of and visualization representations of traceability using graph-based tools.

Definition 1. *A **unified structure**($\mathcal{US}$) is a tuple* $\langle \mathrm{ME}, \prec, \overset{1}{\hookrightarrow}, \cdots, \overset{n}{\hookrightarrow}, \lambda_m, \lambda_d, \overset{1}{\tau}, \cdots, \overset{m}{\tau} \rangle$ *with*

- ME*, a finite set of the model elements,*
- $\prec \subseteq \mathrm{ME} \times \mathrm{ME}$*, the* containment relation *such that it is an (irreflexive) partial order,*
- $\lambda_m \subseteq \mathrm{ME} \times \mathrm{ME}$*, the constraint on model elements,*
- $\lambda_d \subseteq \mathrm{ME} \times (\prec \cup \overset{1}{\hookrightarrow} \cup \cdots \cup \overset{n}{\hookrightarrow})$*, the constraint on dependencies,*
- $\forall i \in \{1, \cdots, n\}, \overset{i}{\hookrightarrow} \subseteq \mathrm{ME} \times \mathrm{ME}$*, the* dependency relation*, and*
- $\forall j \in \{1, \cdots, m\}, \overset{j}{\tau} \subseteq \mathrm{ME}$*, the* type set *of model elements such that* $\forall e \in \mathrm{ME}, \exists \tau \in \{\overset{1}{\tau}, \cdots, \overset{m}{\tau}\} : e \in \tau$.

Here, for all $x, y \in \mathrm{ME}$, $x \overset{i}{\hookrightarrow} y$ ($i \in \{1, \cdots, n\}$) is called a *dependency*, read as x *depending on* y (note that i denotes that the type of dependencies). And $x \prec y$ means x is contained in y. If $x \prec z, y \prec z$, they are simplistically denoted by $x, y \prec z$ and means x and y are both contained in z. For all $w, v \in \mathrm{ME}$, the notation $v \not\prec w$ means that w does not contain v. The tuples $\overset{1}{\tau}, \cdots, \overset{m}{\tau}$ are grouping constructs for model elements and are used to classify the model elements.

We then present an example showing how UML diagrams are converted into $\mathcal{US}$ models. Figure 1 presents UML diagrams of a video-on-demand (VOD) system that allows user U to select or play movies provided by server S. Different from the original one [5], our example adds a *Composite State* in the state machine diagram and the corresponding *Combined Fragment* in the sequence diagram. The structure of the VOD system is represented

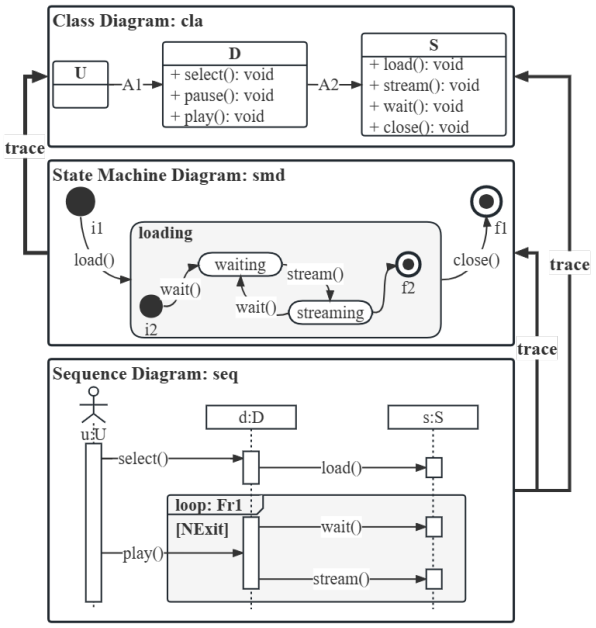


Figure 1. VOD System Example

in the class diagram (top), where the state machine diagram(middle) specifies the behavior of the class S in the class diagram, and the sequence diagram (bottom) depicts watching a movie.

Example 1. *By Definition 1, the entire VOD system is denoted as* $\mathcal{US}_{\mathcal{VOD}} = \langle \mathrm{ME}, \prec, \lambda_m, \lambda_d, \overset{ClaD}{\tau}, \overset{SeqD}{\tau}, \overset{StaD}{\tau}, \overset{Trace}{\hookrightarrow} \rangle$ *where* $\mathrm{ME} = \{seq, smd, cla, (cla, seq), (cla, smd), (smd, seq)\}$, $\prec = \lambda_m = \lambda_d = \emptyset$, $\overset{ClaD}{\tau} = \{cla\}$, $\overset{SeqD}{\tau} = \{seq\}$, $\overset{StaD}{\tau} = \{smd\}$, *and* $\overset{Trace}{\hookrightarrow} = \{(cla, seq), (cla, smd), (smd, seq)\}$. $ClaD$ *refers to class diagrams,* $SeqD$ *means sequence diagrams,* $StaD$ *denotes state machine diagrams, and* $Trace$ *is dependencies between these diagrams. Intuitively, this is constructed at a higher level which views the entire system as a diagram. The US supports not only high-level modeling but also concrete one.*

All UML diagrams in Figure 1 can be independently modeled by Definition 1. The class diagram cla, one of elements in $\overset{ClaD}{\tau}$ *of* $\mathcal{US}_{\mathcal{VOD}}$*(i.e., the* cla *element), is represented by is* $\mathcal{US}_{cla} = \langle \mathrm{ME}', \prec', \lambda'_m, \lambda'_d, \overset{Class}{\tau}, \overset{Oprs}{\tau}, \overset{Asoc}{\tau} \rangle$ *where* $\mathrm{ME}' = \{U, D, S, select, play, load, stream, wait, close\}$, $\prec' = \lambda'_m = \lambda'_d = \emptyset$, $\overset{Class}{\tau} = \{U, D, S\}$, $\overset{Asoc}{\tau} = \{A1, A2\}$, *and* $\overset{Oprs}{\tau} = \{select, play, load, stream, wait, close\}$. *There are notations such as* $Class$ *representing classes,* $Asoc$ *representing associations, and* $Oprs$ *depicting operations of all classes.*

The sequence diagram seq is denoted as $\mathcal{US}_{seq} = \langle$

$\mathrm{ME}'', \prec'', \overset{Inst}{\hookrightarrow}, \overset{Seq}{\hookrightarrow}, \overset{Itrc}{\hookrightarrow}, \lambda''_m, \lambda''_d, \overset{Obj}{\tau}, \overset{Msg}{\tau}, \overset{Frag}{\tau}, \overset{Grd}{\tau}\rangle$ *where* $\mathrm{ME}'' = \{u, s, d, U, S, D, Fr1,\ NExit,\ select,\ load,\ wait, play, stream\}$, $\prec''= \{(NExit, Fr1), (wait, Fr1), (stream, Fr1)\}$, $\lambda''_m = \lambda''_d = \emptyset$, $\overset{Obj}{\tau}= \{u, d, s\}$, $\overset{Frag}{\tau}= \{ Fr1\}$, $\overset{Grd}{\tau} = \{NExit\}$, $\overset{Msg}{\tau}= \{select, load,\ wait,\ play,\ stream\}$, $\overset{Inst}{\hookrightarrow} = \{(d, D), ((s, S), (u, U)\}$, $\overset{Seq}{\hookrightarrow}= \{(load,\ select), (wait, load),\ (play, wait),\ (stream, play)\}$, *and* $\overset{Itrc}{\hookrightarrow}= \{(d, u), (s, d)\}$. $\prec''$ *is non-empty due to combined fragment* $Frag$ *containing* Guards *represented by* Grd. *The notation* $Inst$ *means the relation between objects* Obj *and corresponding classes,* Seq *represents the sequence of all messages* Msg, *and* $Itrc$ *represents the interactions between objects.*

And the state machine diagram smd *can be denoted by* $\mathcal{US}_{smd} = \langle \mathrm{ME}''', \prec''', \overset{Trans}{\hookrightarrow}, \lambda'''_m, \lambda'''_d, \overset{I}{\tau}, \overset{F}{\tau}, \overset{CS}{\tau}, \overset{SS}{\tau}, \overset{Acts}{\tau}\ \rangle$ *where* $\mathrm{ME}''' = \{i2, i1, s2, waiting, streaming, loading, f1, f2, load, wait, stream, close\}$, $\prec'''= \{(i2, loading), (waiting, loading), (streaming, loading), (f2, loading), (wait, loading), (stream, loading)\}$, $\overset{Trans}{\hookrightarrow} = \{(loading, i1), (waiting, i2), (waiting, streaming), (streaming, waiting), (f2, streaming), (f1, loading)\}$, $\lambda'''_m = \lambda'''_d = \emptyset$, $\overset{I}{\tau}= \{i1, i2\}$, $\overset{F}{\tau}= \{f1, f2\}$, $\overset{CS}{\tau}= \{\ loading\ \}$, $\overset{SS}{\tau} = \{ waiting, streaming\}$, *and* $\overset{Acts}{\tau} = \{load,\ wait,\ stream,\ close\}$. *We classify states of the state machine diagram into initial states* I, *final states* F, *simple states* SS, *and composite states* CS. *We named the transitions in the state machine diagram as* $Trans$. *Containment relation* $\prec'''$ *is formed between composite states and the elements inside.*

It is obvious that the unified structure model can easily specify single UML model and the composition of UML models.

Definition 2. *Let* $\mathcal{US} = \langle \mathrm{ME}, \prec, \overset{1}{\hookrightarrow}, \cdots, \overset{n}{\hookrightarrow}, \lambda_m, \lambda_d, \overset{1}{\tau}, \cdots, \overset{m}{\tau}\rangle$ *be a unified structure.*

(1) A sequence $rc = x_1 \cdots x_n$ *is called a* relation chain *in* $\mathcal{US}$ *iff* $\forall i \in \{1, \cdots, n-1\}, x_i, x_{i+1} \in \mathrm{ME}$, $(x_i, x_{i+1}) \in (\prec \cup \overset{1}{\hookrightarrow} \cup \cdots \cup \overset{n}{\hookrightarrow}) \vee (x_{i+1}, x_i) \in (\prec \cup \overset{1}{\hookrightarrow} \cup \cdots \cup \overset{n}{\hookrightarrow})$.

$\ddot{rc}$ *denotes the model elements in the relation chain* $rc = x_1 \cdots x_n$, *that is,* $\ddot{rc} = \{x_1, \cdots, x_n\}$. $RC(\mathcal{US})$ *denotes all possible relation chains in* $\mathcal{US}$.

(2) A sequence $dc = x_1 \cdots x_n$ *is called a* dependency chain *in* $\mathcal{US}$ *iff* $\forall i \in \{1, \cdots, n-1\}, x_i, x_{i+1} \in \mathrm{ME}$, $(x_i, x_{i+1}) \in (\prec \cup \overset{1}{\hookrightarrow} \cup \cdots \cup \overset{n}{\hookrightarrow})$.

$\hat{dc}$ *denotes the model elements in the dependency chain* $dc = x_1 \cdots x_n$, *that is,* $\hat{dc} = \{x_1, \cdots, x_n\}$. $DC(\mathcal{US})$ *denotes all possible dependency chains in* $\mathcal{US}$. $[dc]$ *denotes the number of model elements in the dependency chain* $dc = x_1 \cdots x_n$, *that is,* $[dc] = n$.

Obviously, a relation chain is nondirectional while a dependency chain is directional. For example, $dc_{trace} = cla\ smd\ seq$,$[dc_{trace}] = 3$ in Figure 1 is one of dependency chains and likewise a relation chain.

Proposition 1. *Let* $\mathcal{US}$ *be a unified structure and* $dc = x_1 \cdots x_n \in DC(\mathcal{US})$. *If* $\forall i \in \{1, \cdots, n-1\}, (x_i, x_{i+1}) \in \prec$, *then there does not exist a cycle in* dc.

This proposition states that a dependency chain only containing containment relations does not have a cycle.

Proposition 2. *If* $\mathcal{US}$ *is a unified structure, then* $DC(\mathcal{US}) \subseteq RC(\mathcal{US})$.

Clearly, the number of relation chains is greater than equal to that of dependency chains in a unified structure.

Complex software systems contain many UML diagrams to specify complete information about the systems. Once the whole system model is constructed, the UML diagrams (or their subparts) must be consistent with information. We then discuss the composition of models.

Definition 3. *Let* $\mathcal{US}' = \langle \mathrm{ME}', \prec', \overset{1}{\hookrightarrow}', \cdots, \overset{n}{\hookrightarrow}', \lambda'_m, \lambda'_d, \overset{1}{\tau}', \cdots, \overset{m}{\tau}'\rangle$ *and* $\mathcal{US}'' = \langle \mathrm{ME}'', \prec'', \overset{1}{\hookrightarrow}'', \cdots, \overset{n}{\hookrightarrow}'', \lambda''_m, \lambda''_d, \overset{1}{\tau}'', \cdots, \overset{m}{\tau}''\rangle$ *be two unified structures.*

$\mathcal{US}'$ *is called a* substructure *of* $\mathcal{US}''$, *denoted as* $\mathcal{US}' \sqsubseteq \mathcal{US}''$, *iff* $\mathrm{ME}' \subseteq \mathrm{ME}'', \prec' \subseteq \prec'', \overset{1}{\hookrightarrow}' \subseteq \overset{1}{\hookrightarrow}'', \cdots, \overset{n}{\hookrightarrow}' \subseteq \overset{n}{\hookrightarrow}''$, $\lambda'_m \subseteq \lambda''_m, \lambda'_d \subseteq \lambda''_d$ *and* $\overset{1}{\tau}' \subseteq \overset{1}{\tau}'', \cdots, \overset{m}{\tau}' \subseteq \overset{m}{\tau}''$.

Example 2. *As Example 1 illustrates, we first construct* $\mathcal{US}_{\mathcal{VOD}}$ *and then independently model each diagram in* $\mathcal{US}_{\mathcal{VOD}}$. *Thus, there exists the substructures* $\mathcal{US}_{cla} \sqsubseteq \mathcal{US}_{VOD}, \mathcal{US}_{seq} \sqsubseteq \mathcal{US}_{VOD}$, *and* $\mathcal{US}_{smd} \sqsubseteq \mathcal{US}_{VOD}$.

A substructure is included in the original unified structure, and the unified structure may be separated into multiple substructures.

Definition 4. *Let* $\mathcal{US}' = \langle \mathrm{ME}', \prec', \overset{1}{\hookrightarrow}', \cdots, \overset{n}{\hookrightarrow}', \lambda'_m, \lambda'_d, \overset{1}{\tau}', \cdots, \overset{m}{\tau}'\rangle$ *and* $\mathcal{US}'' = \langle \mathrm{ME}'', \prec'', \overset{1}{\hookrightarrow}'', \cdots, \overset{n}{\hookrightarrow}'', \lambda''_m, \lambda''_d, \overset{1}{\tau}'', \cdots, \overset{m}{\tau}''\rangle$ *be two unified structures.*

If $\prec' \cup \prec''$ *is an (irreflexive) partial order, the composition of* $\mathcal{US}'$ *and* $\mathcal{US}''$ *is defined as* $\mathcal{US}' \uplus \mathcal{US}'' = \langle \mathrm{ME}, \prec, \overset{1}{\hookrightarrow}, \cdots, \overset{n}{\hookrightarrow}, \lambda_m, \lambda_d, \overset{1}{\tau}, \cdots, \overset{m}{\tau}\rangle$ *where* $\mathrm{ME} = \mathrm{ME}' \cup \mathrm{ME}''$, $\prec = \prec' \cup \prec''$, $\forall i \in \{1, \cdots, n\}$: $\overset{i}{\hookrightarrow} = \overset{i}{\hookrightarrow}' \cup \overset{i}{\hookrightarrow}''$, $\lambda_m = \lambda'_m \cup \lambda''_m, \lambda_d = \lambda'_d \cup \lambda''_d$ *and* $\forall j \in \{1, \cdots, m\}$: $\overset{j}{\tau} = \overset{j}{\tau}' \cup \overset{j}{\tau}''$. $\mathcal{US}', \mathcal{US}''$ *are said to be* composable.

Note that two composable unified structures may have different number of types of dependency relations and elements, we equivalently translate them into the two unified

structures with the same number of dependency or element types before composition. For example, $\mathcal{US}'$ and $\mathcal{US}''$ are composable where $\mathcal{US}' = \langle \mathrm{ME}', \prec', \overset{a}{\hookrightarrow}', \overset{x}{\hookrightarrow}', \overset{n}{\hookrightarrow}', \lambda'_m, \lambda'_d, \overset{1}{\tau}', \cdots, \overset{m}{\tau}' \rangle$ and $\mathcal{US}'' = \langle \mathrm{ME}'', \prec'', \overset{x}{\hookrightarrow}'', \overset{y}{\hookrightarrow}'', \overset{z}{\hookrightarrow}'', \lambda''_m, \lambda''_d, \overset{1}{\tau}'', \cdots, \overset{m}{\tau}'' \rangle$. Obviously, we can translate $\mathcal{US}'$ and $\mathcal{US}''$ into $\mathcal{US}_1$ and $\mathcal{US}_2$, respectively:

$$\mathcal{US}_1 = \langle \mathrm{ME}', \prec', \overset{a}{\hookrightarrow}', \overset{x}{\hookrightarrow}', \overset{y}{\hookrightarrow}', \overset{z}{\hookrightarrow}', \lambda'_m, \lambda'_d, \overset{1}{\tau}', \cdots, \overset{m}{\tau}' \rangle$$

where $\overset{y}{\hookrightarrow}' = \overset{z}{\hookrightarrow}' = \emptyset$ and

$$\mathcal{US}_2 = \langle \mathrm{ME}'', \prec'', \overset{a}{\hookrightarrow}'', \overset{x}{\hookrightarrow}'', \overset{y}{\hookrightarrow}'', \overset{z}{\hookrightarrow}'', \lambda''_m, \lambda''_d, \overset{1}{\tau}'', \cdots, \overset{m}{\tau}'' \rangle$$

where $\overset{a}{\hookrightarrow}'' = \emptyset$.

Clearly, $\mathcal{US}_1 = \mathcal{US}', \mathcal{US}_2 = \mathcal{US}''$. Moreover, $\mathcal{US}_1$ and $\mathcal{US}_2$ have the same number of dependency types. Thus, $\mathcal{US}_1$ and $\mathcal{US}_2$ can be composed according to the previous definition. The composition of unified structures has the following properties.

Proposition 3. *Let $\mathcal{US}$, $\mathcal{US}'$ and $\mathcal{US}''$ be three unified structures. And let every two of the three unified structures be composable. Then*

(1) $\mathcal{US}' \uplus \mathcal{US}''$ is a unified structure,

(2) $\mathcal{US}' \uplus \mathcal{US}'' = \mathcal{US}'' \uplus \mathcal{US}'$, and

(3) $(\mathcal{US} \uplus \mathcal{US}') \uplus \mathcal{US}'' = \mathcal{US} \uplus (\mathcal{US}' \uplus \mathcal{US}'')$.

This proposition shows the composition of unified structures has closure, commutativity, and associativity.

Proposition 4. *If $\mathcal{US}$, $\mathcal{US}'$ are two unified structures and composable, then $(DC(\mathcal{US}) \cup DC(\mathcal{US}')) \subseteq DC(\mathcal{US} \uplus \mathcal{US}')$.*

The composition of unified structures does not add or reduce any elements of native structures. Consequently the dependency chain of unified structures remains after composition.

3 Consistency Relation

We shall introduce the atomicity, composition, and equivalence of the consistency relation in this section after providing a definition of the consistency relation. Consistency can be treated as a relation that stores the pairs of either two elements in UML diagrams or two UML models, which satisfies the contained consistency rules.

A comparison should be sought between two or more UML models with UML consistency rules. Consistency rules are systematically collected in plain English text [18]. All external and internal rules are required to ensure the consistency of a system model. Assuming that the UML models of VOD example in Figure 1 are *internally* consistent, we select and simplify a few *external* rules (See Table 1).

Table 1. Several Consistency Rules

ID	Description
CR1	Each class in the class diagram must be instantiated in a sequence diagram.
CR2	Each public method in a class diagram triggers a message in a sequence diagram.
CR3	A message referring to operation must belong to operations of the class that types the lifeline.

Definition 5. *Let $\mathcal{US}' = \langle \mathrm{ME}', \prec', \overset{1}{\hookrightarrow}', \cdots, \overset{n}{\hookrightarrow}', \lambda'_m, \lambda'_d, \overset{1}{\tau}', \cdots, \overset{m}{\tau}' \rangle$ and $\mathcal{US}'' = \langle \mathrm{ME}'', \prec'', \overset{1}{\hookrightarrow}'', \cdots, \overset{n}{\hookrightarrow}'', \lambda''_m, \lambda''_d, \overset{1}{\tau}'', \cdots, \overset{m}{\tau}'' \rangle$ be two unified structures. A relation $R_C \subseteq \mathrm{ME}' \times \mathrm{ME}''$ is called a* consistency relation *between $\mathcal{US}'$ and $\mathcal{US}''$ iff there exists a consistency rule between $\mathcal{US}'$ and $\mathcal{US}''$ such that R_C satisfies the correspondence between $\mathcal{US}'$ and $\mathcal{US}''$ under such a rule. We define $R_C|_{\mathcal{US}'} = \{e \in \mathrm{ME}' \mid \exists e' \in \mathrm{ME}'' : (e, e') \in R_C\}$ and $R_C|_{\mathcal{US}''} = \{e \in \mathrm{ME}'' \mid \exists e' \in \mathrm{ME}' : (e', e) \in R_C\}$.*

Here, we formally denote the consistency rule as a binary relation.

Example 3. *According to the consistency rules CR1 and CR2 in Table 1, there exist the corresponding consistency relations $R_{C1} = \{(D, d), (S, s)\}$ and $R_{C2} = \{(select, select), (play, play), (load, load), (stream, stream), (wait, wait)\}$ between the class diagram $\mathcal{US}_{cla}$ and the sequence diagram $\mathcal{US}_{seq}$ in Figure 1.*

Proposition 5. *Let $\mathcal{US}'$ and $\mathcal{US}''$ be two unified structures. If R_{C1}, R_{C2} be two consistency relations between $\mathcal{US}'$ and $\mathcal{US}''$, then $R_{C1} \cup R_{C2}$ be a consistency relation between $\mathcal{US}'$ and $\mathcal{US}''$.*

Example 4. *As R_{C1}, R_{C2} is presented in Example 3, we have $R_C = R_{C1} \cup R_{C2} = \{(D, d), (S, s), (select, select), (play, play)\}$. Obviously, R_C is the composition of the consistency rules CR1 and CR2.*

Proposition 6. *Let $\mathcal{US}'$ and $\mathcal{US}''$ be two unified structures. And let R_{C1}, R_{C2} be two consistency relations between $\mathcal{US}'$ and $\mathcal{US}''$. Then*

(1) $R_{C1} \cup R_{C2}$ be a consistency relation between $\mathcal{US}'$ and $\mathcal{US}''$.

(2) $R_{C1} \cap R_{C2}$ be a consistency relation between $\mathcal{US}'$ and $\mathcal{US}''$.

(3) $R_{C1} \setminus R_{C2}$ be a consistency relation between $\mathcal{US}'$ and $\mathcal{US}''$.

This proposition states that the consistency relations are preserved under the union, intersection and minus operations

Theorem 1. *Let $\mathcal{US}_1$, $\mathcal{US}'_1$, $\mathcal{US}_2$, $\mathcal{US}'_2$ be four unified structures. Let $\mathcal{US}_1 \sqsubseteq \mathcal{US}'_1$ and $\mathcal{US}_2 \sqsubseteq \mathcal{US}'_2$. If R_C be a consistency relation between $\mathcal{US}_1$ and $\mathcal{US}_2$, then R_C be a consistency relation between $\mathcal{US}'_1$ and $\mathcal{US}'_2$.*

Clearly, the consistency between the submodels implies the consistency between the original models.

To manage consistencies between models, it is necessary to handle how consistencies influence mutually. We introduce *atomic* consistencies, i.e., the consistencies that cannot be further decomposed. The combination of atomic consistencies can express those complicated consistencies. Thus, we discuss how the combination works in the following.

Definition 6. *Let $\mathcal{US}' = \langle \mathrm{ME}', \prec', \overset{1}{\hookrightarrow}', \cdots, \overset{n}{\hookrightarrow}', \lambda'_m, \lambda'_d, \overset{1}{\tau}', \cdots, \overset{m}{\tau}' \rangle$ and $\mathcal{US}'' = \langle \mathrm{ME}'', \prec'', \overset{1}{\hookrightarrow}'', \cdots, \overset{n}{\hookrightarrow}'', \lambda''_m, \lambda''_d, \overset{1}{\tau}'', \cdots, \overset{m}{\tau}'' \rangle$ be two unified structures. Let R_C be a consistency relation between $\mathcal{US}'$ and $\mathcal{US}''$. R_C is said to be* atomic *iff $\exists\ \overset{x}{\tau}' \in \{\overset{1}{\tau}', \cdots, \overset{m}{\tau}'\} : R_C|_{\mathcal{US}'} \subseteq \overset{x}{\tau}'$ and $\exists\ \overset{y}{\tau}'' \in \{\overset{1}{\tau}'', \cdots, \overset{m}{\tau}''\} : R_C|_{\mathcal{US}''} \subseteq \overset{y}{\tau}''$.*

This states that an atomic consistency is only connected to the type sets of model elements when the model is translated into corresponding unified structures.

Example 5. *Figure 1 presents the consistency relation $R_C = \{((d,u), A1), ((s,d), A2)\}$ between $\mathcal{US}_{seq}$ and $\mathcal{US}_{cla}$. Because (d,u) and (s,d) violate $R_C|_{\mathcal{US}_{seq}}$ according to Definition 6, R_C is not* atomic .

Theorem 2. *Let $\mathcal{US}' = \langle \mathrm{ME}', \prec', \overset{1}{\hookrightarrow}', \cdots, \overset{n}{\hookrightarrow}', \lambda'_m, \lambda'_d, \overset{1}{\tau}', \cdots, \overset{m}{\tau}' \rangle$ and $\mathcal{US}'' = \langle \mathrm{ME}'', \prec'', \overset{1}{\hookrightarrow}'', \cdots, \overset{n}{\hookrightarrow}'', \lambda''_m, \lambda''_d, \overset{1}{\tau}'', \cdots, \overset{m}{\tau}'' \rangle$ be two unified structures. Let R_C be a consistency relation between $\mathcal{US}'$ and $\mathcal{US}''$.*

If R_C is not atomic, then there exists the k atomic consistency relations $R_{C_1}, \cdots, R_{C_k}$ such that $R_C = R_{C_1} \cup \cdots \cup R_{C_k}$.

This theorem considers that the atomic consistency relations are the most fundamental parts as they can compose new consistency relations. There is a common mechanism for UML diagrams to construct models and their sub models. For instance, a sequence diagram uses Combined Fragment to contain explicitly a set of interactions, and a class diagram uses relations(e.g., composition, aggregation, realization, and inheritance) to express implicit containment of model elements. Both containments have not yet been universally discussed.

Theorem 3. *Let $\mathcal{US}' = \langle \mathrm{ME}', \prec', \overset{1}{\hookrightarrow}', \cdots, \overset{n}{\hookrightarrow}', \lambda'_m, \lambda'_d, \overset{1}{\tau}', \cdots, \overset{m}{\tau}' \rangle$ and $\mathcal{US}'' = \langle \mathrm{ME}'', \prec'', \overset{1}{\hookrightarrow}'', \cdots, \overset{n}{\hookrightarrow}'', \lambda''_m, \lambda''_d, \overset{1}{\tau}'', \cdots, \overset{m}{\tau}'' \rangle$ be two unified structures. Let R_C be a consistency relation between $\mathcal{US}'$ and $\mathcal{US}''$.*

If $(a,b) \in R_C$ and there exists a relation $R_x \subseteq \mathrm{ME}' \times \mathrm{ME}''$ such that $\forall (e', e'') \in R_x, e' \prec' a \wedge e'' \prec'' b$,then R_x is a consistency relation between $\mathcal{US}'$ and $\mathcal{US}''$.

The theorem shows that the consistency relation is preserved under containment.

Example 6. *There exists the consistency relation $R_C = \{(loading, Fr1)\}$ between $\mathcal{US}_{seq}$ and $\mathcal{US}_{smd}$ in Figure 1.*

In the sequence diagram $\mathcal{US}_{seq}$, there are the messages $wait, stream$ satisfying $wait \prec'' Fr1, stream \prec'' Fr1$. Similarly, the state machine diagram $\mathcal{US}_{smd}$ has two actions $wait, stream$ that satisfy $wait \prec''' loading, stream \prec''' loading$. By Theorem 3, there exists the consistency relation $R_x = \{(wait, wait), (stream, stream)\}$ between $\mathcal{US}_{seq}$ and $\mathcal{US}_{smd}$.

Next, we discuss the composition of consistencies.

Theorem 4. *Let $\mathcal{US}_1$, $\mathcal{US}'_1$, $\mathcal{US}_2$, $\mathcal{US}'_2$ be four unified structures. Let $\mathcal{US}_1, \mathcal{US}'_1$ be composable and $\mathcal{US}_2, \mathcal{US}'_2$ be composable. If R_C be a consistency relation between $\mathcal{US}_1$ and $\mathcal{US}_2$, then*

(1) R_C be a consistency relation between $\mathcal{US}_1 \uplus \mathcal{US}'_1$ and $\mathcal{US}_2 \uplus \mathcal{US}'_2$,

(2) R_C be a consistency relation between $\mathcal{US}_1$ and $\mathcal{US}_2 \uplus \mathcal{US}'_2$, and

(3) R_C be a consistency relation between $\mathcal{US}_1 \uplus \mathcal{US}'_1$ and $\mathcal{US}_2$.

Proposition 3 shows that types of dependency relations and elements during composition between unified structures are incremental. Thus, the consistency relation between models before composition remains.

Example 7. *In Example 3, we have the consistency relation $R_{C1} = \{(D,d), (S,s)\}$ between $\mathcal{US}_{cla}$ and $\mathcal{US}_{seq}$ according to CR1 in Table 1. As is stated by Proposition 3, the composition of unified structures is a unified structure. Let a unified structure $\mathcal{US}' = \mathcal{US}_{seq} \uplus \mathcal{US}_{smd}$. By Definition 4, R_{C1} is consistency relation between $\mathcal{US}_{cla}$ and $\mathcal{US}'$ as well. Analogously, Theorem 4 (3) is shown.*

In our method, the consistency relations are equivalent only if both ends of the consistency relation are model elements. Those non-type sets of the unified structure express implicit and explicit relationships between model elements.

Definition 7. *Let R_{C1}, R_{C2} be two consistency relations. R_{C1}, R_{C2} are* equivalent*, denoted by $R_{C1} \leftrightsquigarrow R_{C2}$ iff both R_{C1} and R_{C2} can be decomposed into the same* atomic *consistency relations.*

The equivalent consistency relations share identical *atomic* consistency relations.

Proposition 7. *Let R_{C1}, R_{C2}, and R_{C3} be three consistency relations.*

(1) $R_{C1} \leftrightsquigarrow R_{C1}$.

(2) $R_{C1} \leftrightsquigarrow R_{C2} \Rightarrow R_{C2} \leftrightsquigarrow R_{C1}$.

(3) $R_{C1} \leftrightsquigarrow R_{C2} \Rightarrow R_{C1} \cup R_{C3} \leftrightsquigarrow R_{C2} \cup R_{C3}$.

This proposition states that the equivalent consistency relations have idempotence and commutativity. Furthermore, the equivalence between the consistency relations preserves if they have the same operations.

4 Tool

We have developed an experimental tool deployed in http://219.151.152.164:3000(See Figure 2). All diagrams are stored in the JSON document form and can be exported in the PNG file form. The tool allows consistency management concerning two main phases: inconsistency detection and inconsistency repair. The outcomes of each detection trigger the generation of repairs for the corresponding model. Moreover, we have implemented the composition of unified structures and equivalence of consistency relations. More functions will be added to the tool step by step.

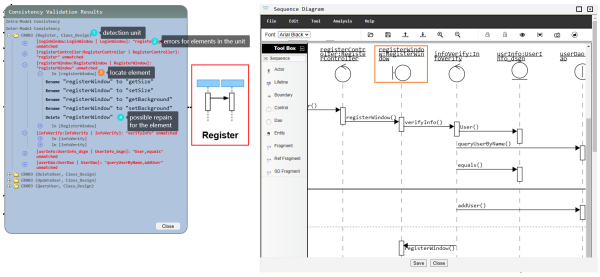


Figure 2. Inconsistencies detection

5 Related Work

A precise semantics for overall UML diagrams have drawn great attentions. Most of the efforts focus on formalizing commonly used UML diagrams. For instance, Lu et al. [12] analyse the behavior aspects of the UML sequence diagram by a trace of properties to check and reason consistency. Based on the transformation into Labeled Transition Systems (LTS), Lambolais et.al [10] propose a framework to combine refinement and extension developments, and then they check for consistency in the incremental process using the accept set (as a special failure trace semantics [16]). The semantic models mentioned above are behavior ones with not good semantics. Instead, we present a new formal model called unified structure which can characterize not only all the UML behavioral diagrams, but also all the UML structural diagrams.

Consistency checking is promising research work in UML, and the techniques are widely discussed. Egyed [7] proposes an informal way of using rule instances to check consistency instantly and efficiently. He further discusses an incremental method to perform consistency checking in [6]. Formal approaches to check consistency are extensively reviewed. For example, Xu et al. [21] check the context's incremental consistency when constraints are represented using a First Order Logic formula, which addresses stopping the system's aberrant behavior. Based on the transformation of UML class diagram detailed in Object Constraint Language (OCL) into a Constraint Satisfaction Problem, Cabot et al. [3] resort the automatic solver to check properties. OCL extends UML diagrams by textual constraints but it limits the expressiveness of UML. And Campbell et al. [4] proposed an intermediate step for integrating UML diagrams into a formal framework to recognize discrepancies between the system's structure and behavior. Therefore, transforming UML diagrams into formal semantics differentiates among existing literature, which leads to different consistency checking methods.

Motivated by Egyed [7, 6], we apply consistency rules for efficient and accurate inconsistency detection. The work of collecting UML consistency rules is ongoing, which comes from the work of Torre et al [18]. A total of 116 consistency rules were systematically documented among 10 of the 14 types of UML diagrams. The network of consistencies brings the complexity of formal verification. Klare and Heiko [9] decompose consistency relations on model transformation by extracting a tree from a reduced consistency relation graph. Our strategy, in contrast, is to allow the decomposition of consistency relations between particular models. This decomposition offers flexibility to depict consistencies. We further introduce a consistency equivalence for simplification and reduction.

6 Conclusion

In this paper, we have introduced a new formal model called unified structure. It represents UML diagrams and enables tooling and analysis for consistencies between UML diagrams. Based on the unified structure model, we have discussed the composition, decomposition, and equivalence of consistency between UML models. This paper aims to define formal consistency relations between UML diagrams. Such consideration provides a foundation for conflict checking of consistency rules.

Threats to validity come from two main aspects. On the one hand, it is vital to observe that constraints and models in our case are limited. On the other hand, consistency relations require the professional to comprehend so that the relations can be used correctly.

In future work, we will first improve scalability of the tool to analyze UML consistency relations and their equiv-

alence. Then we will conduct experiments in complicated scenarios to update and optimize our tool.

References

[1] J. Abualdenien and A. Borrmann. A meta-model approach for formal specification and consistent management of multi-LOD building models. *Adv. Eng. Informatics*, 40:135–153, 2019.

[2] M. N. Alanazi and D. A. Gustafson. Super state analysis for uml state diagrams. In *WRI CSIE*, volume 7, pages 560–565. IEEE, 2009.

[3] J. Cabot, R. Clarisó, and D. Riera. On the verification of UML/OCL class diagrams using constraint programming. *JSS*, 93:1–23, 2014.

[4] L. A. Campbell, B. H. C. Cheng, W. E. McUmber, and R. E. K. Stirewalt. Automatically detecting and visualising errors in UML diagrams. *Requirements Engineering*, 7(4):264–287, 2002.

[5] A. Egyed. Instant consistency checking for the uml. In *ICSE*, pages 381–390, 2006.

[6] A. Egyed. UML/Analyzer: A Tool for the Instant Consistency Checking of UML Models. In *ICSE*, pages 793–796. IEEE, 2007.

[7] A. Egyed. Automatically detecting and tracking inconsistencies in software design models. *IEEE TSE*, 37(2):188–204, 2010.

[8] Y. Hammal. A modular state exploration and compatibility checking of uml dynamic diagrams. In *AICCSA*, pages 793–800. IEEE, 2008.

[9] H. Klare. Multi-model consistency preservation. In *MoDELS*, pages 156–161. ACM, 2018.

[10] T. Lambolais, A.-L. Courbis, H.-V. Luong, and C. Percebois. IDF: A framework for the incremental development and conformance verification of UML active primitive components. *JSS*, 113:275–295, 2016.

[11] K. Lano. Formal specification using interaction diagrams. In *SEFM*, pages 293–304. IEEE, 2007.

[12] L. Lu and D.-K. Kim. Required behavior of sequence diagrams. *ACM TOSEM*, 23(2):1–28, 2014.

[13] D. A. Meedeniya, I. D., and I. Perera. Software Artefacts Consistency Management towards Continuous Integration: A Roadmap. *IJACSA*, 10(4), 2019.

[14] F. u. Muram, H. Tran, and U. Zdun. Systematic review of software behavioral model consistency checking. *CSUR*, 50(2):1–39, 2017.

[15] OMG. Unified Modeling Language, v2.5.1. https://www.omg.org/spec/UML/2.5.1/PDF. [Online; accessed 2017-12].

[16] H. Ponce de León, S. Haar, and D. Longuet. Conformance relations for labeled event structures. In *Tests and Proofs: 6th International Conference*, pages 83–98. Springer, 2012.

[17] P. Stünkel, H. König, Y. Lamo, and A. Rutle. Comprehensive Systems: A formal foundation for Multi-Model Consistency Management. *FAC*, 33(6):1067–1114, Dec. 2021.

[18] D. Torre, Y. Labiche, M. Genero, and M. Elaasar. A systematic identification of consistency rules for UML diagrams. *JSS*, 144:121–142, oct 2018.

[19] D. Torre, Y. Labiche, M. Genero, and et al. Uml consistency rules: a case study with open-source uml models. In *SEFM*, pages 130–140, 2020.

[20] H. Wen, J. Wu, J. Jiang, G. Tang, and Z. Hong. A Formal Approach for Consistency Management in UML Model. *IJSEKE*, 2023.

[21] C. Xu, S. C. Cheung, and W. K. Chan. Incremental consistency checking for pervasive context. In *ICSE*, pages 292–301. ACM, 2006.

[22] J. Yang, Q. Long, Z. Liu, and X. Li. A predicative semantic model for integrating uml models. In *ICTAC*, pages 170–186. Springer, 2005.

[23] X. Zhao, Q. Long, and Z. Qiu. Model checking dynamic uml consistency. In *ICFEM*, pages 440–459. Springer, 2006.

Building and Assessing an Italian Textual Dataset for Emotion Recognition in Human-Robot Interactions

Alessia Fantini[1,5], Antonino Asta[2], Alfredo Cuzzocrea[3,4] *, Giovanni Pilato[5]
[1]University of Pisa, Pisa, Italy
[2]University of Palermo, Palermo, Italy
[3]iDEA Lab, University of Calabria, Rende, Italy
[4]Dept. of Computer Science, University of Paris City, Paris, France
[5]ICAR-CNR, Italian National Research Council, Palermo, Italy
alessia.fantini@icar.cnr.it, antonino.asta@community.unipa.it
alfredo.cuzzocrea@unical.it, giovanni.pilato@icar.cnr.it

Abstract

In this study, we illustrate an ongoing work regarding building an Italian textual dataset for emotion recognition for HRI. The idea is to build a dataset with a well-defined methodology based on creating ad-hoc dialogues from scratch. Once that the criteria had been defined, we used ChatGPT to help us generate dialogues. Human experts in psychology have revised each dialogue. In particular, we analyzed the generated dialogues to observe the balance of the dataset under different parameters. During the analysis, we calculated the distribution of context types, gender, consistency between context and emotion, and interaction quality. With "quality" we mean the adherence of text to the desired manifestation of emotions. After the analysis, the dialogues were modified to bring out specific emotions in specific contexts. Significant results emerged that allowed us to reorient the generation of subsequent dialogues. This preliminary study allowed us to draw lines to guide subsequent and more substantial dataset creation in order to achieve increasingly realistic interactions in HRI scenarios.

1 Introduction

Emotions are key factors during Human-Robot Interaction (HRI). At the same time, one of the most difficult tasks for robots during interaction with humans is emotion recognition[21], [12]. Emotions have a multidimensional nature and their understanding depends on the context in which they are expressed. Context is a key element in understanding of emotions and one of the challenges in NLP research. Context makes it possible to predict emotion to some degree. For example, being at a party, finding a new job, taking a trip with very high probability are related to the emotion of "joy". Similarly, a bereavement or an argument with a loved one tends to be associated with "sadness".
It is clear that emotions can overlap, they can be different from person to person, and the same context can generate one emotion at one time and a different emotion at another time, but we tend to be able to identify objective situations to which specific emotions are linked. So providing examples of context-related emotions can help in this regard. In [22], talking about conversational context modeling, the authors state that context can make it possible to significantly improve the NLP systems. Within data-driven models, therefore, it is critical to build a dataset that is as specific and contexual as possible.

There are many contributions in the literature regarding the construction of datasets for emotion recognition. Most of them cover few emotions, tending only to Ekman's basic ones. Some examples are EmotionX [26] , Affect-Intensity Lexicon and Emotion Dataset (AILA) [18], CrowdFlower's Emotion Dataset [1], Friends [14] , EmoBank [6]. Furtheremore, many approaches build dataset using news paper, books or dialogues found on the Internet, including those found from social media, e.g. SemEval-2018 Task 1: Affect in Tweets (AIT-2018) [19], Sentiment140 [13], Emotion Intensity Dataset (EmoInt) [17], The International Survey on Emotion Antecedents and Reactions (ISEAR) [24]. Others use movies, e.g. The Stanford Sentiment and Emotion Classification (SSEC) [25, 20] or physiological signals, e.g. The DEAP (Database for Emotion Analysis using Physiological Signals) [15].

*This research has been made in the context of the Excellence Chair in Big Data Management and Analytics at UPC, Paris, France

DOI reference number: 10.18293/DMSVIVA2023-006

Regarding Italian dataset, there are fewer contributions and often from tweets, some of the most widely used include SEMEVAL-ITA-2018 [7], ITA-EVALITA-2020 [3], EmoLexIta [9], The STS-ITA (Sentences in the Wild - Italian) [5], or news articles e.g. News-ITA [23]. A lexicon based approach has been also used for sentiment classification of books reviews in the Italian language [8].

With respect to the main contributions from the literature, we decided to avoid data from social media or newspaper articles as these have specific language that sometimes does not fit well with natural interactions. For usage scenarios such as ours, thus that of Human-Robot Interactions, we decided to use examples of interactions between people in which the emotions we want to focus on. This is an important feature of our study, since thanks to the dialogue structure it is possible to provide the robot with examples of interactions very similar to those that occur in the real world. By creating *ad hoc* dialogues, therefore, we could also provide the specific context in which certain emotions may emerge. Also, the labeling was not done directly by us: this is another of the challenges highlighted by [22] in conversational context. We asked the ChatGPT to generate dialogues in which a specific emotion, such as *joy*, emerges; subsequently, we monitored and possibly adjusted or validated the associated labeling. Another important point of our study is that we not only include basic emotions, but we label a total of fourteen emotions by assuming those that may possibly emerge during HRI in contexts such as home, medical, school, but also in everyday life. These emotions are *joy, sadness, anger, fear, surprise, disgust, frustration, embarrassment, boredom, nervousness, melancholy, guilt, hope, and stress*. Finally, according to our perspective, a good emotional dataset should have a balance in the data from different perspectives. In order to achieve this goal, we performed a further analysis on the dialogues generated by exploiting ChatGPT, calculating different quantities, such as the distribution of gender, the type of context, the consistency between emotion and context, and, in general, we evaluated the quality of the interaction.

The remainder of the paper is organized as follows: the next section illustrates the methodology that we used to build the dataset, then a sample of the collected and modified dialogues as well as the subsequent analysis is reported; then in section 4 a brief discussion is given about the dataset characteristic; in the end conclusions and future work are illustrated.

2 Methodology

Our work aims to build an Italian dataset for dialog-based emotion recognition. To generate dialogues, we first defined methodological criteria, and then we exploited ChatGPT to help us develop them by taking advantage of the speed in data generation. Once the dialogues were generated, human psychology experts reviewed each conversation to analyze the adequacy of the dataset from different points of view. We analyzed consistency between requested emotion and context, gender distribution, types of context generated, and quality of interaction, understood as the appropriateness of language concerning specific emotions. The methodology comprises three stages: dialogue generation procedure, data analysis, and improvements.

2.1 Procedure

For each emotion (14 in total), we decided to generate 25 dialogues. The command given to ChatGPT was to generate a short conversation, of about five lines, between two people in which a specific emotion emerges. Next, we decided to generate five dialogues for each emotion by asking ChatGPT not to use the word corresponding to the emotion, and we labeled these kind of dialogues "Without Word (W.W.)". This was done to test whether ChatGPT could generate discussions in which, e.g., sadness emerged without having the word "sadness" in the text. The goal is to create data that increasingly reflect real situations to train robots that can recognize emotions based on context and not just by recognizing specific words. The small number is because this is a pilot study to build a more extensive dataset later. Finally, the original dialogues generated were retained, but we created a copy to edit them after performing the analysis. Both the Web interface and the API provided by OpenAI were used. This has made it possible to obtain different styles of narrations of the events. Gpt 3.5-turbo model was used, with the following role: "You are a writer assistant who produces dialogue that accurately reflects emotion".

2.2 Analysis

Dialogues were analyzed considering four factors: consistency between context and emotion, gender distribution, type of contexts, and quality of interaction. By **consistency (C)** between context and emotion, we mean whether the context generated is consistent with the feeling expressed. For example, the context of an argument with the boss is a context compatible with the emotion of anger. So for each dialogue, we assessed whether or not there was consistency. We counted the percent relative frequency.

$$C = \frac{Nyes}{Ndialogues} \cdot 100$$

Similarly, for **gender distribution(GD)**, we counted how many times the gender "Neutral (*N*), Masculine (*M*) and Feminine (*F*)" occurred in the dialogues and we calculated the percent relative frequency.

$$GD = \frac{Ngender(NorMorF)}{Ntotgender} \cdot 100$$

Regarding the **type of context(TC)**, we created classes and counted how many belonged to each class; then, we calculated the percent relative frequency.

$$TC = \frac{NcontextX}{Ntotcontexts} \cdot 100$$

The classes identified are *Work*, *Leisure*, *Luck*, *Interpersonal sphere*, *Generic*. In some cases, we identified a specific category, e.g., in the "Disgust" dialogues, we identified the category "Animals and Objects," as several scenarios expressed disgust for objects or animals.

Finally, for the **quality of interaction(QoI)**, we analyzed the appropriateness of language in expressing a specific emotion. This was evaluated with three values: "*Sufficient*", "*Not much*", "*No*". By "Sufficient (S)" we mean that the language appears natural enough and reflects in the terms used the emotion. By "Not much (NM)" we mean that the language is not very natural and it does not entirely reflect the emotion, e.g., using words that also represent other emotions, but all in all, it is acceptable. By "No (N)," we mean confusion, unusual terms, and/or language that does not reflect the specific emotion. Also, for this parameter, we calculated the percent relative frequency.

$$QoI = \frac{NValue(SorNMorN)}{Ntotinteractions} \cdot 100$$

2.3 Improvements

After the analysis, we conducted several modifications, both grammatically and in terms of content. Another important aspect was observing the distribution of the type of contexts and selecting those most inherent to interpersonal and social scenarios for inclusion in the dataset we will build after this pilot study. To obtain various scenarios, first, it was asked to generate five possible social scenarios in which a specific emotion can emerge. In this way, it was possible to select those scenarios that were more consistent with HRI, or once an interesting one is generated; it was asked to modify it in order to focus on social interaction. Then for each of these scenarios was asked to create a dialogue and then, if necessary, to expand it. Often the model failed to expand the dialogue without the recurring use of the emotion terms, so it was asked to replace them with some expressions that could be metaphors or equivalent expressions. When asked to change scenarios, some emotions were confused. For example, when the emotion of anger was requested, the dialogues generated expressed the emotion of frustration, often repeating the term "frustrating" in the text and vice versa. Similarly, it happened for stress and nervousness. So for these emotions that could generate confusion, it was first asked to provide a definition that clearly distinguished the two emotions. For example, it was asked to provide a definition that clearly distinguishes between frustration and anger. Then based on the definition, it was asked to generate scenarios in which emotion could emerge distinctly. Actually, the scenarios developed were more specific, distinguishing the two emotions. The same was done for stress and nervousness. This demonstrates the importance of the human expert intervening in all phases to direct ChatGPT to generate more focused dialogues.

3 Results

The results will be shown first according to a global view and then in detail for each emotion.

3.1 Global Analysis

With respect to **consistency**, 86% of the generated contexts are consistent with emotion. An example of consistency is this:

- Person 1: Ciao, come stai oggi? (*Hello, how are you today?*)
- Person 2: Non molto bene, sinceramente. (*Not very well, honestly.*)
- Person 1: Oh, mi dispiace sentirlo. C'è qualcosa che ti preoccupa? *(Oh, I'm sorry to hear that. Is something bothering you?*)
- Person 2: Sì, ho avuto un brutto litigio con un caro amico ieri sera. Siamo finiti per dire cose terribili l'un l'altro e ora mi sento come se avessi perso una parte di me stesso. (*Yes, I had a bad fight with a close friend last night. We ended up saying terrible things to each other and now I I feel like I've lost a part of myself.*)
- Person 1: Capisco. Litigi del genere sono sempre molto difficili da gestire. Ma forse potresti cercare di parlare con il tuo amico e chiarire la situazione? (*I understand. Fights like that are always very difficult to deal with. But maybe you could try to talk to your friend and clear the air?*)
- Person 2: Ho provato a chiamarlo stamattina, ma non ha risposto. Non so se vuole ancora parlarmi. (*I tried to call him this morning, but he didn't answer. I don't know if he still wants to talk to me.*)
- Person 1: Mi dispiace tanto. Spero che alla fine riuscirete a risolvere le cose. Se hai bisogno di parlare, sai che puoi contare su di me. (*I'm so sorry. I hope you will eventually be able to work things out. If you need to talk, you know you can count on me.*)

- Person 2: Grazie, apprezzo molto il tuo sostegno. Ma per ora mi sento solo triste e confuso. (*Thank you, I really appreciate your support. But for now I just feeling sad and confused.*)

In this dialogue, the context is an argument with a friend, and effectively the sadness that emerges is consistent with the type of context. In some cases, the generated contexts did not reflect the corresponding emotion. An example:

- Person 1: Ciao, come stai? (*Hello, how are you?*)
- Person 2: Non molto bene, devo dire. (*Not very well, I must say.*)
- Person 1: Che cosa è successo? (*What happened?*)
- Person 2: Stamattina mi sono svegliato con un forte mal di stomaco. Ho passato tutto il mattino in bagno. (*I woke up this morning with a bad stomachache. I spent the whole morning in the bathroom.)*
- Person 1: Ah, mi dispiace sentirlo. Cosa pensi possa essere? (*Ah, I'm sorry to hear that. What do you think it might be?)*
- Person 2: Non ne sono sicuro, ma credo sia stato qualcosa che ho mangiato ieri sera. Ho provato un nuovo ristorante messicano, ma credo che il cibo non fosse fresco. (*I'm not sure, but I think it was something I ate last night. I tried a new Mexican restaurant, but I think the food wasn't fresh.*)
- Person 1: Oh no, che schifo. Ti senti meglio adesso? (*Oh no, that's gross. Do you feel better now?*)
- Person 2: Un po' meglio, ma non ancora del tutto. Devo ancora prendere qualche medicina. (*A little better, but not quite yet. I still have to take some medicine.*)
- Person 1: Sì, è meglio prendersi cura di te stesso. Spero tu ti riprenda presto. (*Yes, you better take care of yourself. I hope you recover soon.*)

In this case, except for "that's gross", the entire dialogue is centered on a context showing a reaction to physical discomfort. The person feels vomiting because of physical discomfort, as he has probably eaten bad food, and therefore not because of an emotional reaction of disgust.

Regarding **gender**, there is a strong imbalance in the dataset: N=30% M=56% F=14%

In addition, in a couple of cases, the gender count was canceled because the same person was first male and then female. Here is an example of a dialogue about frustration:

- Person 1: Ho lavorato duramente su questo progetto ma non ha (grammar error) ottenuto il successo sperato. (*I worked hard on this project but it did not (in the Italian version-grammar error) achieve the success I had hoped for.*)
- Person 2: Mi dispiace sentirti così deluso (indicates that person 1 is male). Cosa pensi sia andato storto? (*I'm sorry to feel so disappointed (in the Italian version indicates that person 1 is male). What do you think went wrong?*)
- Person 1: Non ne sono sicuro, ho messo tutta me stessa (female gender) ma sembra che non sia abbastanza. (*I'm not sure, I put all of myself (in the Italian version-female gender) but it seems like it's not enough.*)
- Person 2: Non scoraggiarti, ogni esperienza è una lezione imparata. Magari hai bisogno di un po' di tempo per riflettere e riprovarci con un approccio diverso. (*Don't be discouraged, every experience is a lesson learned. Maybe you need some time to reflect and try again with a different approach.*)

The **context** overall appears heterogeneous but it is unbalanced when observed in relation to specific emotions. For example, for the emotion "Joy," only three types of context were generated. Specifically, ten contexts are about *success* (e.g., passing a university exam, promotion at work), ten are about *leisure* (e.g., traveling, starting a yoga class), four are about *luck* (e.g., winning the lottery), and only one is about *Personl life situations* (receiving a gift). The type of context will be discussed in depth in the description of each emotion.

Regarding the **quality of interaction**, the adherence of text to the desired manifestation of emotions was evaluated. In 65% of the dialogues, we can define the quality of interaction as "sufficient". However, some changes were added later either in terms of grammatical corrections or to make the dialogue more fluid and natural. In 25% of cases, there is a poor fit between text and emotion. Finally, in 10% of the dialogues, the text was completely garbled or did not reflect the desired emotion. Here are some examples of the three categories:

Sufficient: Boredom

- Friend 1: "Cosa c'è che non va, sembri distratta?" (*"What's wrong, you seem distracted?"*)
- Friend 2: "Sì, sto solo pensando ad altro. Questa lezione mi fa venire la noia." (*"Yes, I'm just thinking about something else. This class is making me bored."*)
- Friend 1: "Capisco come ti senti, anche io sto trovando difficoltà a restare concentrata." (*"I understand how you feel, I am also finding it hard to stay focused."*)

- Friend 2: ”Sì, vorrei solo essere altrove ora. Anche voi pensate la stessa cosa, giusto?” (*”Yes, I just wish I was somewhere else right now. You also feel the same way, right?”*)

- Friend 1: ”Sì, credo che tutti vorremmo fare altro.” (*”Yes, I think we all would like to do something else.”*)

Not much: Anger

- Person 1: ”Non ci posso credere, ho perso l’autobus per la terza volta questa settimana!” (*“I can’t believe I missed the bus for the third time this week!”*)

- Person 2: ”Ma come hai fatto?” (*”But how did you do it?”*)

- Person 1: ”Non mi hai visto? Mi hai tenuto a parlare e l’autobus è passato sotto il mio naso!” (*”Didn’t you see me? You kept me talking and the bus passed right under my nose!”*)

- Person 2: ”Non è colpa mia se sei sempre in ritardo!” (*”It’s not my fault you’re always late!”*)

- Person 1: ”Ma certo che è colpa tua! Non riesci mai a smettere di parlare e poi ti lamenti se arrivo sempre tardi!” (*”Of course it’s your fault! You can never stop talking and then you complain that I’m always late!”*)

- Person 2: ”Ok, ok, calmati! Non c’è bisogno di arrabbiarsi!” (*”Okay, okay, calm down! No need to get angry!”*)

- Person 1: ”Ma come faccio a non arrabbiarmi? Questo mi fa perdere tempo e soldi!” (*”But how can I not get angry? This wastes my time and money!”*)

- Person 2: ”Hai ragione, mi dispiace. Cercherò di essere più attento la prossima volta.” (*”You’re right, I’m sorry. I’ll try to be more careful next time.”*)

No: Hope

- Character A: ”Spero solo di non sembrare troppo stressato/a stasera.” (*”I just hope I don’t look too stressed out tonight.”*)

- Character B: ”Non preoccuparti, sei bellissimo/a e la serata sarà fantastica.” (*”Don’t worry, you look beautiful and the evening will be great.”*)

- Character A: ”Speriamo che ci siano delle sorprese piacevoli stasera, vorrei che fosse tutto diverso dal solito.” (*”Hopefully there will be some pleasant surprises tonight, I’d like everything to be different than usual.”*)

- Character B: ”Stasera sarà diversa dal solito, perché sarà proprio come ci piace. Semplice e piena di speranze!” (*”Tonight will be different than usual, because it will be just the way we like it. Simple and hopeful!”*)

3.2 Single Emotion Analysis

Below we show the analysis of each of the 14 emotions according to the 4 parameters outlined in the methodology section.

- **JOY**
  - Consistency = 100%
  - Gender = N 12% M 80% F 8%
  - Contexts = 10 Success, 10 Leisure, 4 Luck, and only 1 is about personal life situations
  - Quality of interaction = Sufficient 64% Not much 36%

- **SADNESS**
  - Consistency = 92%
  - Gender = N 46% M 54% F 0
  - Contexts = Heterogeneous mainly generic and interpersonal
  - Quality of interaction = Sufficient 88% Not much 12%

- **ANGER**
  - Consistency = 100%
  - Gender = N 12% M 55% F 33%
  - Contexts = Heterogeneous, sometimes reactions out of proportion to the context
  - Quality of interaction = Sufficient 88% Not much 12%

- **FEAR**
  - Consistency = 100%
  - Gender = N 24% M 72% F 4%
  - Contexts = Mostly related to horror contexts (shadows, animals, running away from someone) - Absence of contexts related to more interpersonal or social fear, such as fear of the future.
  - Quality of interaction = Satisfactory 72% Not much 28%

- **SURPRISE**
  - Consistency = 100%

- Gender = N 24% M 76% F 0%
- Contexts = Heterogeneous
- Quality of interaction = Satisfactory 88% Not much 12%

- **DISGUST**
  - Consistency = 96%
  - Gender = N 72% M 28% F0
  - Contexts = Highly related to foods, insects, objects. No examples related to people's behaviors or abstract concepts. Only in two cases is there a reference to disgust as a result of a person's behavior.
  - Quality of interaction = Sufficient 84% Not much 16%

- **FRUSTRATION**
  - Consistency = 28%: in three cases there is confusion with *anger*
  - Gender = N 46% M 50% F 4%
  - Contexts = Heterogeneous, sometimes reactions out of proportion to the context
  - Quality of interaction = Sufficient 80% Not much 20%

- **EMBARRASSMENT**
  - Consistency = 68% sometimes there is confusion with *guilt*.
  - Gender = N 44% M 48% F 8%
  - Contexts = Heterogeneous
  - Quality of interaction = Sufficient 76% Not much 24%

- **BOREDOM**
  - Consistency = 92%
  - Gender = N 16% M 56% F 28%
  - Contexts = Heterogeneous, mainly leisure time
  - Quality of interaction = Sufficient 56% Not much 28% No 16%

- **NERVOUSNESS**
  - Consistency = 88%
  - Gender = N0 M 53% F 47%
  - Contexts = Heterogeneous
  - Quality of interaction = Sufficient 68% Not much 20% No 12%

- **MELANCHOLY**
  - Consistency = 88%
  - Gender = N 40% M 60% F0
  - Contexts = Heterogeneous
  - Quality of interaction = Sufficient 68% Not much 16% No 16%

- **GUILT**
  - Consistency = 92%
  - Gender = N 40% M 40% F 20%
  - Contexts = 24% relate to work contexts, while most are related to interpersonal or social situations (e.g., arguing with a friend, neglecting family, telling a lie, etc...)
  - Quality of interaction = Satisfactory 52% Not much 44% No 4% . In many dialogues the language appears out of proportion to the emotion

- **HOPE**
  - Consistency = 100%
  - Gender = N 52% M 36% F 16%
  - Contexts = 28% relate to work contexts, 44% relate to medical contexts, 28% relate to interpersonal or social situations
  - Quality of interaction = Satisfactory 28% Not much 64% No 8% . Often the language seems to belong more to fear or nervousness and not to hope. Here is an example:

- Studente 1: "Sto preparando questo esame da giorni, spero di ottenere un buon voto." (*"I've been preparing for this exam for days, I hope to get a good grade."*)
- Studente 2: "Sono sicuro che andrà tutto bene, hai studiato tanto e sai quello che fai." (*"I'm sure you'll do well, you've studied hard and you know what you're doing."*)
- Studente 1: "Sì, ma ho paura di non ricordare tutte le informazioni durante l'esame." (*"Yes, but I'm afraid I won't remember all the information during the exam."*)
- Studente 2: "Non preoccuparti, vai tranquillo e non lasciare che l'ansia ti prenda il sopravvento. Spero che otterrai la valutazione che meriti." (*"Don't worry, go easy and don't let anxiety get the best of you. I hope you will get the grade you deserve."*)
- Studente 1: "Grazie per il supporto! Ho davvero bisogno di sentirlo. Speriamo che andrà tutto bene." (*"Thank you for the support! I really need to hear it. Hopefully everything will be okay."*)

- Studente 2: "Sarà così, cerca di rilassarti e di essere positivo. La speranza è la chiave del successo." (*"It will, just try to relax and be positive. Hope is the key to success."*)
- **STRESS**
  - Consistency = 84% . Sometimes confusion with nervousness
  - Gender = N 0 M 83% F 17%
  - Contexts = 56% relate to work contexts, the remaining are heterogeneous
  - Quality of interaction = Satisfactory 60% Not much 12% No 28%

3.3 Analysis of Dialogues "Without Words"

As for dialogues generated without the word expressing the emotion emerging, most generated contexts **consistent** with the emotion. Not very consistent were the dialogues of Boredom, Fear, Surprise and Hope. In contrast, only 1 out of 5 of the Disgust dialogues was found to be consistent. The other 4 were so confused that they could not be classified. Here is an example:

- Person 1: "Che schifo...quella lì sembra venuta direttamente dal nido delle rane!" (*"Gross...that one looks like it came straight from the frogs' nest!"*)
- Person 2: "Davvero, mi ha fatto venire il voltastomaco. Però non possiamo farci niente al riguardo... forse conviene che andiamo a prendere un po' d'aria fresca." (*"Really, it made me sick to my stomach. We can't do anything about it though...maybe we'd better go get some fresh air."*)
- Person 1: "Sì, direi che hai ragione. Quando si vedono situazioni simili, l'unica cosa da fare è portare il nonno in braccio in giro per la città, almeno la compagnia è più gradevole!" (*"Yes, I would say you are right. When you see situations like that, the only thing to do is to carry Grandpa around town, at least the company is more pleasant!"*)

Furthermore, right among the dialogues of Disgust-W.W. a stereotype about Neapolitans emerged. Here is the dialogue:

- Person 1: "Oh, guardate quella lì...è più volgare di un bidone di spazzatura a Napoli!" (*"Oh, look at that one...she's grosser than a garbage can in Naples!"*)
- Person 2: "Sì, mi ha messo i brividi appena l'ho vista. Comunque, perché non ci beviamo un po' d'acqua insieme? Così ci togliamo lo schifo di bocca!" (*"Yeah, she gave me the creeps as soon as I saw her. Anyway, why don't we have some water together? That way we can get the filth out of our mouths!"*)
- Person 1: "Mi pare un'ottima idea, non vedo l'ora di liberarmi di questa sensazione." (*"That sounds like a great idea, I can't wait to get rid of this feeling."*)

It is not only not at all sufficient from the point of view of language, but a stereotype clearly emerges. Regarding **gender** and **contexts**, the number of dialogues is small to draw specific inferences, however, we can say that they seem to reflect the general trend. As for the **quality of interaction**, it appears worse than the basic dialogues, that is, the non-W.W. dialogues. In fact, in 43% of the cases the quality of interaction was rated as "sufficient", in 32% of the cases "not very much", and in 25% "no". The sum of "not very much" and "no" is also 57% thus exceeding the percentage of those considered sufficient.

4 Discussion

The dataset analysis identified strengths and weaknesses that will allow for guidelines for constructing the larger dataset. Consistency between context and emotion is the main strength as it allows for high reliability in automatically generating dialogues: this allows for fast data generation. Clearly, as the results show, dialogues must always be validated by a human operator, as the conversations generated were not always consistent. Also, concerning the type of contexts, we need to be careful so that they are as heterogeneous as possible, perhaps by including contexts increasingly inherent in the interpersonal and social spheres that reflect possible HRI situations. The analysis of the gender distribution allowed us to observe the large imbalance in favor of the male gender. This will enable us to correct a bias and reflect more generally on training AI systems that need to be as heterogeneous as possible. Regarding this point, the case of dialogue in which there is a stereotype about the city of Naples should also give us some thought. Finally, the quality of the interaction almost always needs modification by the human operator, either for grammatical errors that are occasionally observed, to adjust the language to the emotion, or to make the dialogues more natural.

5 Conclusions and Future work

We conducted a pilot study to guide the construction of an Italian dataset for emotion recognition. After determining the methodology and defining the procedure, we used ChatGPT to generate dialogues quickly. Together with professionals specialized in psychology, we analyzed 420 dialogues about 14 emotions to check the balance of the dataset

from different points of view (context-emotion consistency, gender distribution, types of context generated, and quality of interaction). The results show that there are advantages and limitations to using automatic dialog generation systems and that, certainly, the construction of the dataset cannot disregard the human operator's control. The most significant advantage is the speed of data generation, and it was seen that, in most cases, there is consistency between emotion and generated contexts. Of course, one still needs to control the dialogues to make the contexts heterogeneous and more focused on interpersonal and social aspects. The study also drew attention to the distribution of gender, which is largely unbalanced on the masculine and therefore will allow later to generate dialogues in which it is explicitly requested that the feminine and neutral genders emerge in a way that balances the dataset. Also, concerning the language used and thus the quality of interaction, numerous changes have been made to the dialogues in terms of grammatical, form, and content corrections. Despite this, however, another advantage was that dialogues could be created from scratch, directing ChatGPT to generate dialogues oriented according to criteria defined a priori by the authors. Future work will exploit the information from this study to create a larger, balanced, HRI-oriented dataset. On the other hand, we aim at integrating our framework with emerging challenges due to novel *big data trends* (e.g., [4, 11, 2, 16, 10]).

References

[1] Crowdflower. 2016. the emotion in text. https://www.figure-eight.com/data/ sentiment-analysis-emotion-text/.

[2] P. P. F. Balbin, J. C. R. Barker, C. K. Leung, M. Tran, R. P. Wall, and A. Cuzzocrea. Predictive analytics on open big data for supporting smart transportation services. *Procedia Computer Science*, 176:3009–3018, 2020.

[3] V. Basile, D. M. Maria, C. Danilo, L. C. Passaro, et al. Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and sspeech tools for italian. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, pages 1–7. CEUR-ws, 2020.

[4] L. Bellatreche, A. Cuzzocrea, and S. Benkrid. *F&A*: A methodology for effectively and efficiently designing parallel relational data warehouses on heterogenous database clusters. In *Data Warehousing and Knowledge Discovery, 12th International Conference, DAWAK 2010, Bilbao, Spain, August/September 2010. Proceedings*, volume 6263 of *Lecture Notes in Computer Science*, pages 89–104. Springer, 2010.

[5] M. Braunhofer, M. Elahi, and F. Ricci. *User Personality and the New User Problem in a Context-Aware Point of Interest Recommender System*, pages 537–549. 01 2015.

[6] S. Buechel and U. Hahn. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain, Apr. 2017. Association for Computational Linguistics.

[7] T. Caselli, N. Novielli, V. Patti, and P. Rosso. Sixth evaluation campaign of natural language processing and speech tools for italian: Final workshop (evalita 2018). In *EVALITA 2018*. CEUR Workshop Proceedings (CEUR-WS. org), 2018.

[8] F. Chiavetta, G. L. Bosco, G. Pilato, et al. A lexicon-based approach for sentiment classification of amazon books reviews in italian language. *WEBIST (2)*, 2016:159–170, 2016.

[9] M. Corazza, S. Menini, E. Cabrio, S. Tonelli, and S. Villata. A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–22, 2020.

[10] A. Coronato and A. Cuzzocrea. An innovative risk assessment methodology for medical information systems. *IEEE Trans. Knowl. Data Eng.*, 34(7):3095–3110, 2022.

[11] A. Cuzzocrea, F. Martinelli, F. Mercaldo, and G. V. Vercelli. Tor traffic analysis and detection via machine learning techniques. In *IEEE BigData, 2017*, pages 4474–4480. IEEE Computer Society, 2017.

[12] A. Cuzzocrea and G. Pilato. A composite framework for supporting user emotion detection based on intelligent taxonomy handling. *Logic Journal of the IGPL*, 29(2):207–219, 2021.

[13] A. Goel, J. Gautam, and S. Kumar. Real time sentiment analysis of tweets using naive bayes. In *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, pages 257–261. IEEE, 2016.

[14] A. Joshi, V. Tripathi, P. Bhattacharyya, and M. J. Carman. Harnessing sequence labeling for sarcasm detection in dialogue from TV series 'Friends'. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 146–155, Berlin, Germany, Aug. 2016. Association for Computational Linguistics.

[15] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.

[16] C. K. Leung, A. Cuzzocrea, J. J. Mai, D. Deng, and F. Jiang. Personalized deepinf: Enhanced social influence prediction with deep learning and transfer learning. In *IEEE BigData, 2019*, pages 2871–2880. IEEE, 2019.

[17] S. Mohammad and F. Bravo-Marquez. WASSA-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–49, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics.

[18] S. M. Mohammad. Word affect intensities. *arXiv preprint arXiv:1704.08798*, 2017.

[19] S. M. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko. Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA, 2018.

[20] S. M. Mohammad, P. Sobhani, and S. Kiritchenko. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–23, 2017.

[21] G. Pilato and E. D'Avanzo. Data-driven social mood analysis through the conceptualization of emotional fingerprints. *Procedia computer science*, 123:360–365, 2018.

[22] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953, 2019.

[23] F. Rollo, G. Bonisoli, and L. Po. Supervised and unsupervised categorization of an imbalanced italian crime news dataset. In *Information Technology for Management: Business and Social Issues: 16th Conference, ISM 2021, and FedCSIS-AIST 2021 Track, Held as Part of FedCSIS 2021, Virtual Event, September 2–5, 2021, Extended and Revised Selected Papers*, pages 117–139. Springer, 2022.

[24] K. R. Scherer and H. G. Wallbott. " evidence for universality and cultural variation of differential emotion response patterning": Correction. 1994.

[25] H. Schuff, J. Barnes, J. Mohme, S. Padó, and R. Klinger. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23, 2017.

[26] B. Shmueli and L.-W. Ku. Socialnlp emotionx 2019 challenge overview: Predicting emotions in spoken dialogues and chats. *arXiv preprint arXiv:1909.07734*, 2019.

Providing Accessible and Supportive User Experience through conversational UI and digital humans

Elena Molinari
TUI Musement
Milan, Italy
elena.molinari@tui.com

Andrea Molinari
Dept. of Industrial Engineering
University of Trento
Trento, Italy
andrea.molinari@unitn.it

Abstract— **Designing E-Health services that are accessible, engaging, and provide valuable information to patients is an endeavor that requires research and validation with potential users. The information needs to be perceived as trustworthy and reliable, in order to promote people's ability to make informed decisions about their health. This article focuses on understanding the potential of conversational user interfaces featuring digital humans as communication agents to provide healthcare-related information to users. The main insights inform whether this interaction style can provide a higher level of accessibility and engagement for users, thus creating a better user experience. Since digital humans are not yet extensively adopted in the healthcare domain, few design guidelines are available. The work followed the human-centered design approach to gather requirements and feedback from users. This led to defining six guidelines and an extensive set of observations about user experience and accessibility.**

Keywords- e-Health, accessibility, user experience, digital humans, conversational user interfaces

I. INTRODUCTION

Digital healthcare services are steadily developing and growing, with an increasing number of people relying on them to manage and monitor their health. However, these services often show low adoption rates and fail to meet their goals. One of the main causes of low adoption rates is the failure to meet patients and healthcare professionals' needs and expectations, due to a lack of understanding of requirements from designers and developers [1]. Other factors that hinder the adoption of e-Health services include usability issues, privacy concerns, culture, and flow disruption [2]. The involvement of patients in the design of e-Health services allows creating a better user experience and is crucial to achieve acceptability and adoption.

When a user is engaging with healthcare systems or services, the most appropriate term is "patient experience", rather than "user experience". The Beryl Institute defines PX as "the sum of all interactions shaped by an organization's culture that influence patient perceptions across the continuum of care" [3]. Changing the term from "user" to "patient" allows considering healthcare-specific concepts, such as health literacy. Research has shown that a positive patient experience is one of the strongest indicators of patient retention and adherence to therapy [4]. For this reason, patient experience becomes one of the main measures of the quality of healthcare systems and products.

e-Health services can make healthcare more accessible to everyone, and thus serve the needs of both patients and healthcare professionals and providers. To achieve this, it is crucial to provide information that is reliable and accurate to the patients. In recent years, the trend of patients turning to other patients and to the internet to find health-related information has seen an increase [5]. Despite the benefits that finding comfort and empathy from others in similar situations can bring, it is crucial to ensure that people have access to reliable and scientific information, and that they are redirected to a professional whenever it is needed.

The research took place within a specific case study proposed by Roche. The company wanted to redesign an informational ophthalmology website to feature a digital human substituting the traditional text-based website. The goal of the redesign was to make the content accessible and available to all users, regardless of their visual acuity level. The scope of the website is to provide informative material about eye conditions, how to recognize them and how to act accordingly.

Low vision and vision impairments is a global health concern: the World Health Organization [6] reports that over 2.2 billion people are living with a form of visual impairment. Additionally, Tham et al. [7] explain that ophthalmology is one of the medical fields that is most lagging in terms of digitalization and e-Health services. This generally led to the ophthalmology sector not being ready nor able to face the COVID-19 crisis. The authors believe that there is potential to develop a more digital approach to ophthalmology. e-Health services aim at extending the scope of healthcare provision [8]. To achieve this, it is crucial to provide information that is reliable and accurate to the patients.

This research activity was conducted as Master thesis work of the main author, and extensive information can be found in the original document submitted at Aalto University [9]. focused on the following research problem: e-Health services aiming at providing valuable and reliable information to potential patients often fail to be emotionally supportive and informative for people who are starting to explore the implications of health conditions.

DOI Reference: 10.18293/DMSVIVA2023-321

Following this, two research questions were formulated to guide the research.

- RQ1: Based on empirical research using human-centred design methods, can conversational user interfaces featuring digital humans help make e-Health services more accessible for patients with vision defects, to provide them with the information they need and with emotional support?
- RQ2: What guidelines can be suggested to foster the improvement of the accessibility and emotional support of e-Health services through conversational interaction?

Emotional support is a crucial aspect to consider when designing e-Health services. In fact, medical interventions are most likely to be successful when the doctors are emotionally supportive and friendly, and when they treat the patient as a peer [10]. The literature shows that, if properly designed and implemented, the natural language-based interaction can increase engagement and lead to improved patient experience [11]. However, it is important to note that conversational agents can set higher expectations from the users due to their realistic nature, which can also result in higher levels of frustration if these expectations are not met [12].

User research with people living with low vision and their caregivers was conducted to understand whether using conversational user interfaces featuring digital humans can provide a more accessible interaction modality. A list of guidelines and best practices was created to help designers approach conversational user interfaces featuring digital humans.

II. METHODOLOGY

A. User groups

Courage and Baxter [13] define three groups of users, based on the impact that the product has on them. The first group, primary users, interacts with the service, and benefits directly from the interaction. Secondary users might not interact with the service themselves, but nonetheless benefit from primary users' interactions, or they can influence these interactions. Finally, tertiary users are people or organizations who have decision power on whether to start using a service, and thus indirectly benefit from its usage. Figure 1 shows the division of the user groups that were considered in this case study.

Primary users include people who are starting to experience a decline in their vision and want to gather information about eye conditions and the caregivers of people living with these conditions. Another potential primary user group is composed by people who have not yet started experiencing vision decline, who nonetheless heard about it and want to gather more information.

Secondary users are healthcare professionals, who might benefit from the service because it relieves them from some work burden, as it can act as a first informative encounter for patients.

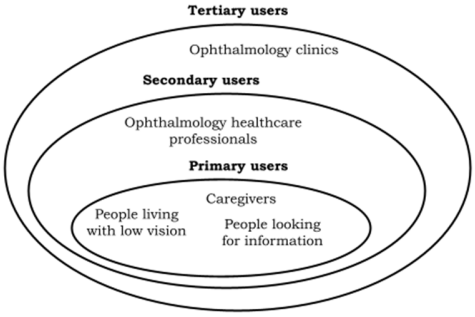


Figure 1 User groups

Tertiary users include ophthalmology clinics, which will benefit from their healthcare professionals having more quality time to dedicate to patients.

B. Research activities focus

The research evaluated several aspects of the CUI-based interface approach. The overarching goal was to ensure that the digital human can be a valuable agent to con-vey health-related information. It evaluated the usability of the conversational user interface. In order to focus the scope of the research, the ergonomic criteria presented by Bastien and Scapin [14] were adopted. In particular, the research focused on the criteria of guidance, workload, explicit control, error management and consistency. The research additionally focused on the accessibility of the conversational user inter-face. Finally, the research investigated the digital human's ability to enhance the emotional engagement and support of users receiving information about eye conditions.

In order to provide a positive experience and a pleasant interaction, the digital human should be perceived as empathic, engaging, and trustworthy. In fact, the relationship between the digital human and the user should be based on trust, to ensure acceptance of the information and the ability to act on it.

A good entry point to building trust in the digital human is having affinity between the agent and the user. Affinity with a digital agent is influenced by the perceived realism of the interaction. A project featuring digital human agents must focus on creating an experience that gives the users the impression that they are interacting with a real person [15]. To achieve this, the interaction between the person and the agent must not trigger the so-called uncanny valley effect, a negative feeling of eeriness and discomfort when interacting with human-resembling characters [16]. The uncanny valley effect theory explains that increased realism and anthropomorphism increase the affinity level that a person perceives for a digital agent. However, there is a point (the uncanny valley), where the resemblance to a real human is very high, but not high enough to be pleasant. This causes feelings of eeriness and discomfort in the users, which then leads to unpleasant experiences with the agent. Movement reinforces this effect: moving stimuli cause a much stronger effect compared to still ones [16].

This effect could strongly impact the users' perceptions of the trustworthiness and reliability of the service. The

overarching goal of the research presented in this article was to ensure that the digital human can be a valuable agent for conveying health-related information.

C. Methodology and participants

Figure 2 shows the interface presented to the participants. The digital human occupies the middle of the screen. On the top left corner there is an accessibility menu, on the bottom left corner a backward button and, on the bottom right corner, a forward button. The white band in the middle shows visual feedback of the captured words when the user is speaking. Next to the forward button, the options among which the users can choose are displayed. The buttons are in grey for non-available content and in orange for available content.

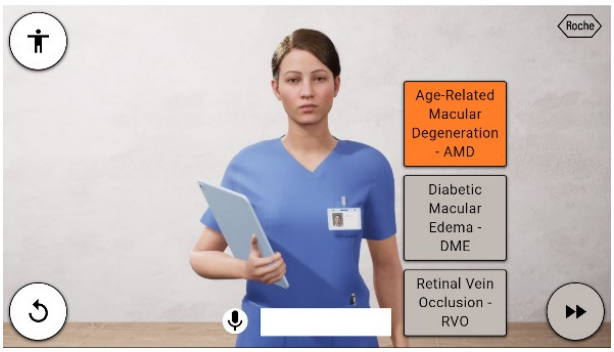


Figure 2 Interface

The interface that the participants interacted with was a high-fidelity horizontal prototype. This means that almost all the functionalities that the final product should contain were developed to a certain extent:

- It was possible to use speech interaction, but only to choose among the presented options, and not to freely interact with the digital human.
- Out of the three conditions that were meant to be addressed in the service, only one flow was working.
- The accessibility options could be explored, but only the text size control worked.

This resulted in the need to give a quite strict and detailed scenario to the participants, for them not to incur in dead ends or errors and to avoid unnecessary frustration.

The chosen research methodology was participant observation, paired with follow-up semi-structured interviews. This methodology is well-suited for the use case because it allows observing how users would naturally interact with the system on the first time that they access it. This article will focus on the user research results, which informed the six guidelines' design.

A total of eighteen participants was recruited. Fourteen of them were recruited through Roche's Patient Advisory board, and ten of them took part in one-on-one interviews. Four more participants were recruited through usertesting.com. Twelve people were living with low vision; six were caregivers or people involved in prevention and advocacy for low vision conditions. Six people were not visually impaired, six were mildly impaired, and six were severely impaired, meaning they could see very little, and three used screen readers.

67% of the participants were older than 50 years of age. This distribution represents the potential target group for an informative ophthalmology service. However, most of the participants were quite knowledgeable about eye conditions, which is not expected from potential service users. Besides this, the target group of this service would generally not be highly visually impaired but looking for more information about eye conditions.

None of the participants had used the service before, thus allowing to investigate discoverability and learnability. This also resulted in the participants having a homogeneous familiarity level with the service, making results more consistent.

The results presented in this article derive from the combination of the fourteen one-on-one calls evaluating the prototype with potential users and two focus groups organized afterwards. The one-on-one calls panel featured ten Patient Advisory board members and four people recruited through usertesting.com. The four usertesting.com participants did not join the focus groups, but four more Patient Advisory board members were involved in this phase. This means that a total of fourteen participants were recruited for both sessions. The focus groups were organized in two rounds: six participants took part in the first focus group and eight participants in the second.

III. Results

The sessions produced sixty-two items (individual observations and comments), both negative and positive, which exemplify the perceptions of the participants' panel. These items were then aggregated and will be presented here in the form of insights, divided in the different aspects that were taken into consideration. These aspects are:

- Usability and guidance: the overall appreciation level of the functionalities of the service and its ability to instruct the users on how to navigate it.
- Accessibility: with a focus on low-vision participants.
- Content, trustworthiness and reliability: quality of the presented content and people's level of trust for the information coming from the digital agent. This aspect is particularly important given the nature of the information provided.
- Emotional support: people's feeling of investment towards the digital human's speech, and the perceived support from the agent.
- Perception of realism: people's perception of the realism of the interaction. This is linked to the Uncanny Valley effect, which might hinder a positive experience with the digital human.

The participants were given a clear scenario and the unavailable content was marked. They were asked to gather more information about a specific eye condition, and they were free to go through the content in the way that they preferred

(interacting through speech or using buttons). The digital human told some information at every step and then allowed the user to select a path by presenting options.

A. Usability and guidance

Two participants reported that the current guidance may not be sufficient for people looking for generic information, meaning users who are not knowledgeable about eye conditions. For this reason, several participants encouraged allowing users to start the navigation from the symptoms. People's personal situations and experiences should be a focus when creating an e-Health service.

The participants noted that the available options and the digital human's speech must not create dissonance. This means that they should be consistent and not confusing. An unclear situation was when the digital human mentioned the eye conditions' names before allowing the users to select one of them. This situation was specifically problematic for visually impaired participants using screen readers, who could not read the options labels and solely rely on the spoken guidance. The research session showed the need for careful design of the options presented.

Three participants remarked that the list of options available at each stage of the interaction was very long. The number of options and their quite complex names make it hard to remember the label for the one to choose. The spoken text should therefore be as short as possible when presenting options. Besides, options should be selectable by number or synchronously with the speech. This means that, when the options are presented, the interface should respond to speech commands such as "this one" or "the first option". Finally, options must be available for restatement at any time.

About 20% of the participants questioned whether the user group of people older than seventy years would be able to use the service, meaning whether they could operate the website and know how to interact with the digital human. This user group could make up a big portion of the intended users, since many eye conditions onset in late life stages. More research would be needed to explore this topic, involving less educated and technically skilled participants.

B. Accessibility

Many of the accessibility issues that were found came from people who were using screen readers or struggled to see the screen in any other way.

As discussed in the previous section, screen reader users struggled the most with presenting options. However, the combination of the digital human's speech and the screen reader listing the different buttons on the screen seemed to work fine. Screen reader users need the assistive technology and appreciate it, so it is not advisable to disable it or force the users not to use it.

One participant said: *"I like my screen reader because I'm now used to the way it speaks. At the beginning I was skeptical about this service for this reason. But after trying it, I have to say I really enjoyed it and I would prefer it over the screen reader"*. Despite this very positive feedback, there was no clear consensus across participants about the added value of the digital human over the screen readers. Some participants said they would prefer to get information from a plain text website using their screen reader because it might be more efficient, while others appreciated the interaction with the digital human.

The research session showed that the possibility of pausing the conversation is a fundamental feature of this kind of service. Some participants reported that they might want to take notes or talk to someone else, but without pausing, they could not do it because they would miss content. Besides this, some participants reported that the service should provide the possibility to enable and disable the speech recognition.

Participants appreciated the flexibility provided by having both speech interaction and buttons. Three participants reported that providing users with as many options as possible is crucial to cater to different preferences and needs.

C. Content, trustworthiness and reliability

In general, the digital human seemed to inspire the impressions and feelings that the team envisioned. Participants perceived the digital human as trustworthy, knowledgeable, invested, and friendly. This suggests that people are likely to build a positive relationship with the digital human, and that they would trust it. In fact, the participants reported that they consider the information reliable.

Over 75% of the participants were highly educated in the field of eye conditions. For this reason, they reported that the level of depth of the content would not be very suitable for them, but that it would have been perfect for someone who is just starting to approach eye conditions for the first time. Almost all participants reported that the service should provide the possibility to go into detail to cater to users with different knowledge levels. For example, they wished to receive more daily life coping suggestions.

D. Emotional support and engagement

About 70% of the participants reported that they found the conversation engaging. One participant said that had there been more content, they "would have liked to explore more". This was not the only positive content in this sense, with another person saying they were *"kind of hooked into it"*.

About 20% of the participants explicitly reported that the DH version of an informative service would be *"less impersonal"* than plain written text. Some participants mentioned that it looked like the digital human was invested and interested in what they were saying.

However, the two focus groups gave very different and contrasting results regarding emotional support, making it hard to draw overarching conclusions. The general trend seems to be that the digital human can provide a generic form of emotional support better than plain text but that the core of the emotional support work should be left to real humans. One participant explained that it would be beneficial to have testimonials from other patients, but that this cannot be provided by the DH, which needs to *"step aside"* and leave space for videos of real people to convey this kind of information.

E. Perception of realism (Uncanny Valley effect)

Most participants (about 90%) considered the voice realistic enough to be both engaging and informative. Sighted participants reported that the digital human looked realistic, up to the point that one participant said that the digital human *"made her feel like she was real"*. Two participants said that they did not like the experience, but there seemed to be no evidence of a strong uncanny valley effect among the participants. The movements of digital humans are crucial to determine perceived realism but only one participant reported that they were distracted by the lack of syncing of the speech and the lips movements. Besides this, some participants did not like some of the movements that the DH was performing, especially while waiting (such as looking at an imaginary watch), because they felt like they were not respectful. These comments show the importance of carefully designing the movements of a digital human.

F. General insights

The results of the user research are encouraging to keep investigating the potential of digital humans and speech interaction to provide information about ophthalmic conditions. However, more work needs to be done to provide a completely accessible service that would answer the needs of all users. The main accomplishment of the case study service is that people consider it a reliable source of information that they would trust and listen to. This is a great achievement that contributes to moving towards the goal of informing people and fostering prevention.

IV. Discussion

A. Answers to RQ1

The results of the research suggest that people tend to react positively to a digital human conveying information about eye conditions. This type of agents seems to have the potential to provide more personalized explanations, which results in a deeper connection with the user. Users reported feeling *"as if someone was there with them"*. The concept was generally well accepted, despite the idea that this kind of interaction might not be for everyone.

People with high visual impairment generally appreciated being talked to rather than going through a static webpage. Caregivers also reported that it could have been beneficial to have this kind of service when their loved ones started to experience symptoms. A conversational user interface embodied by a digital human can provide personalization and a human touch, which people appreciate.

For people with a low-severity condition, speech interaction allows avoiding the fatigue from reading content on a screen. People living with a high-severity condition enjoy hearing a human-like voice instead of the more robotic screen reader's voice (albeit the latter being much faster and possibly more efficient). Furthermore, providing content optimized for listening and not for reading is an advantage because written content follows a generally more complex structure than spoken one.

In conclusion, conversational user interfaces using digital humans as communication agents appear to have great potential in providing users (especially low-vision users) with healthcare-related information. This information is perceived as factual and trustworthy, and the additional support that a conversational agent can provide is appreciated and can contribute to a better user experience.

B. Answers to RQ2

The research insights allowed building a set of six guidelines that designers and researchers should consider when creating a conversational user interface-based service. These guidelines are shown in Table 1.

TABLE I. Resulting guidelines

Code	Guideline
G1	Ensure that the digital human is as realistic as possible, not only in its looks, but also in its movements.
G2	Create a conversation flow that is clear and easy to follow. Do not use long sentences and reduce the language complexity as much as possible. Focus on UX writing.
G3	When presenting options, do it in the simplest and most rapid way possible, and allow users to listen to the options as many times as they want. Allow flexibility on how the options can be chosen.
G4	Ensure that a text version of the content is also available.
G5	Ensure that navigation is easy and as self-explanatory as possible.
G6	Ensure compatibility with assistive technologies and provide flexibility, personalization and integration.

1) Guideline 1

To provide a good user experience and smooth interaction, the digital human's appearance must be realistic in looks and movements. Movement appears to be a determinant for the perception of realism and therefore needs to be both fluid and plausible. The design of the digital human should not only focus on the fluidity of movements (which is utterly important), but also on their plausibility. This is well documented by Mori's work [14] and was observed during the tests.

One practical example from the case study is that when the digital human was waiting for the user to make a choice, she would scratch her head or look at her watch. This is something that no real human would ever do because it can easily be perceived as rude or inappropriate. It then results in a lower score in terms of realism.

2) Guideline 2

The content must be optimized for speech. The ways written and spoken content are structured are very different. For this reason, it is important to focus efforts on user experience writing to provide a conversational flow that feels natural to the listener, easy to understand, and provides all the necessary information.

Optimising content for speech allows to avoid flow disruptions, one of the main reasons for missed adoption and dropout of e-Health services [2].

3) Guideline 3

One factor that seems to play a major role in the perception of CUI is how options are presented and selected. Choosing among options must be as simple and as straightforward as possible to achieve a successful interaction. Choosing entails a high cognitive load for the receiver, who needs to remember all the options and then go through the decision process.

In spoken interaction, losing track of the different options and forgetting them is easy. The digital human must also be able to repeat the options as often as needed. Finally, it is extremely important to provide flexibility in choosing options. The conversational user interface should be trained to recognize synonyms, cardinal indications, partial answers, and answers synched to the speech. Implementing a good mechanism for choosing options allows to act on the ergonomic criteria of guidance, workload, explicit control [14]. This allows for much more natural interaction and, thus a better user experience.

4) Guideline 4

The service must provide a written version of the content as well as a spoken one. This can benefit users with hearing impairment, users who do not appreciate conversational user interfaces, and users who are in a hurry or already know what content they want to look for. Providing a text-based version of the content would greatly enhance the user-friendliness of the website, and it is a standard accessibility practice.

5) Guideline 5

One area of concern that needs to be addressed is the navigation of the service, especially when a lot of content organized on different topics is added to the information structure. It is worth considering whether it would be beneficial to have a menu, using standard navigation within the CUI.

Otherwise, navigation possibilities should be provided in another form, for example by requesting the DH to navigate to a different section. The ability to pause, go backward, and skip forwards in the digital human's speech is fundamental to providing a positive user experience because it allows flexibility. Speech commands should be intuitive and easy to trigger. This once again improves the flow of the interaction with the system, which is crucial to increase adoption and improve the user experience [2].

6) Guideline 6

Screen readers should not be disabled or discouraged during the interaction. For this reason, compatibility between the conversational user interface and assistive technologies should be the goal instead of complete substitution. The digital human is, in this case, an improvement of the user experience in that it makes the user feel like they are interacting with a more human entity, rather than to a digital system, which improves engagement [11].

However, the need for assistive technologies should be reduced as much as possible by providing self-explanatory ways to navigate and interact with the interface. The digital human's speech and the screen reader output should not be antagonizing one another but working synergically to provide the best user experience possible. This requires testing with users who regularly utilize screen readers to navigate digital services.

C. Limitations

The research has limitations that need to be acknowledged, and that can inform the planning of future research. In summary, the main methodological limitations are:

- The small number of participants in the user-based sessions.
- Their rather homogeneous demographics and background.
- The little availability of testable content, which impacts the ability to test how people would navigate the content and whether the information architecture could support meaningful exploration.
- The inability to run complete and thorough tests with a consistent group of primary users with lower technological skills, lower or no knowledge about eye conditions and starting to experience vision loss.

A qualitative study like the one performed in this case study relies on a small amount of in-depth data coming from individuals rather than a big sample of quantitative data. For this reason, future work should focus on incorporating quantitative analysis to complement the insights coming from a qualitative-based approach.

Acting on these limitations would provide more generalizable results, allowing for an experience that caters to all users' needs. Future research should be conducted to ensure that the concept of receiving healthcare-related information is well accepted by people who are not very skilled with digital services. Testing whether different cultural backgrounds or different age groups show different opinions about the experience could also provide valuable insights.

Future research is also needed to gauge the limits of the potential of CUIs and digital humans. Having a clear overview of the areas where the digital human cannot provide a positive user and patient experience is important for the development of similar services.

In general, more research is needed to be able to confidently affirm that CUIs using a digital human as a conversational agent are a good tool to provide healthcare-related information. Nonetheless, the results that have been presented in this article are encouraging and show the potential of such solutions.

V. CONCLUSIONS

This work focused on the case study suggested by Roche of re-imagining an ophthalmology patient's website to leverage a conversational user interface approach. The goal was to evaluate whether using a digital human as a conversational agent would provide a better user experience and higher emotional support to users looking for information about eye conditions. The case study allowed for a broader discussion about the potential that digital humans have in offering healthcare-related information. Understanding whether people would consider the information

coming from such an agent to be trustworthy and reliable is crucial for the success of this kind of services.

The research provides insights into the positive and negative aspects of having digital humans as agents in a conversational user interface. Users generally appreciated the concept: they found it engaging, trustworthy and easy to use. However, there are some aspects that could not be addressed during this research, and which need further understanding.

The primary areas that need to be addressed are guidance, navigation, and error management. Nonetheless, the positive feedback gathered from the participants of the evaluation sessions indicates that it is worth investing in the research and development of this relatively new services. In fact, the work showed that conversational user interfaces and digital humans have the potential to positively impact the user experience of informational websites providing healthcare-related content, both in terms of accessibility and engagement.

The six guidelines that resulted from the research activity give initial directions for designers and developers to build conversational user interface featuring digital humans. However, they will need to be complemented with other guidelines emerging from further research.

Future research should be conducted to ensure that the concept of receiving healthcare-related information is well accepted by people who are not very skilled with digital services. In fact, conversational user interfaces might have the potential to make websites more accessible for people who generally struggle with technology and the Internet, but this needs to be checked systematically, to provide a generalizable result. Besides this, conducting more quantitative research might be valuable to ensure that the appreciation of the interaction with digital humans can be proved through statistical evidence as well.

More research can also ensure the generalizability of the results, but the current outlook is positive. In fact, the results showed that the current level of technological ability to reproduce a human generally manages to provide a positive experience for users interacting with the agent. Besides this, people find the information trustworthy and reliable, which is crucial when conveying healthcare-related information.

Based on the results and insights collected through this research, e-Health services can leverage the capabilities of conversational user interfaces and digital humans to provide a better user experience.

ACKNOWLEDGMENTS

We would like to thank the supervisors for this work, Loïc Martinez Normand (PhD), from the Universidad Politécnica de Madrid, and Johanna Kaipio, from Aalto University. We would then like to thank the team at Roche Finland, especially Hanna Helotera, Damien Vincent Lockner and Christine Moriceau.

REFERENCES

[1] Å. Cajander, C. Grünloh, T. Lind, and I. Scandurra, "Designing eHealth services for patients and relatives," Proceedings of the 9th Nordic Conference on Human-Computer Interaction. doi:10.1145/2971485.2987670J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73, 2016.

[2] C. Granja, W. Janssen, and M. A. Johansen. "Factors Determining the Success and Failure of eHealth Interventions: Systematic Review of the Literature". Journal of medical Internet research, 20(5), e10235. https://doi.org/10.2196/10235. 2018.

[3] The Beryl Institute. Patient experience 101 - why? Retrieved March 16, 2022, from https://www.theberylinstitute.org/page/PX101_Why. 2022

[4] M. Bahja and M.Lycett. "Identifying patient experience from online resources via sentiment analysis and topic modelling". Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies. doi:10.1145/3006299.3006335. 2016.

[5] P. Briggs, C. Hardy, P. Harris, and E. Silence. " Patient-led perspectives on ehealth: How might hyperpersonal data inform design?" Proceedings of HCI KOREA 2015 (HCIK '15), 115–121. 2015.

[6] Word Health Organization, "Blindness and vision impairment," retrieved March 16, 2022 from: https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment, 2021.

[7] Y.-C. Tham, R. Husain, K. Y. Teo, A. C. Tan, A. C. Chew, D. S. Ting, C.-Y. Cheng, G. S. Tan, and T. Y. Wong. "New digital models of care in ophthalmology, during and beyond the COVID-19 pandemic," British Journal of Ophthalmology, 106(4), 452–457. https://doi.org/10.1136/bjophthalmol-2020-317683, 2021.

[8] G. Eysenbach, "What is e-health?," Journal of medical Internet research, 3(2), E20. https://doi.org/10.2196/jmir.3.2.e20, 2001.

[9] E. Molinari. "Leveraging Conversational User Interfaces and Digital Humans to Provide an Accessible and Supportive User Experience on an Ophthalmology Service". M.S. thesis. SCI Dept. Aalto University. Espoo. 2022.

[10] P. J. Moore, A. E. Sickel, J. Malat, D. Williams, J. Jackson, and N. E. Adler, "Psychosocial factors in medical and psychological treatment avoidance: The role of the doctor–patient relationship," Journal of Health Psychology, 9(3), 421-433, 2004

[11] R. Kocielnik, R. Langevin, J. S. George, S. Akenaga, A. Wang, D. P. Jones, A. Argyle, C. Fockele, L. Anderson, D. T. Hsieh, K. Yadav, H. Duber, G. Hsieh, and A. L. Hartzler. " Can I talk to you about your social needs? understanding preference for conversational user interface in health". CUI 2021 - 3rd Conference on conversational user interfaces. https://doi.org/10.1145/3469595.3469599. 2021.

[12] L. Ciechanowski, A. Przegalinska, M. Magnuski, and P. GloorIn the shades of the Uncanny Valley: An experimental study of human–chatbot interaction. Future Generation Computer Systems, 92, 539–548. https://doi.org/10.1016/j.future.2018.01.055. 2019

[13] C. Courage, and K. Baxter. "Understanding Your Users" 10.1016/B978-1-55860-935-8.X5029-5. (2005).

[14] J. M. C. Bastien, and D. L. Scapin, "A validation of ergonomic criteria for the evaluation of human-computer interfaces," International Journal of Human-Computer Interaction, 4(2), 183 - 196. https://doi.org/10.1080/10447319209526035, 1992.

[15] M. Seymour, L.I. Yuan, A. Dennis, and K. Riemer. "Have We Crossed the Uncanny Valley? Understanding Affinity, Trustworthiness, and Preference for Realistic Digital Humans in Immersive Environments." Journal of the Association for Information Systems, 22(3), 9. (2021).

[16] M. Mori, K. F. MacDorman, and N. Kageki, "The uncanny valley [from the field]" IEEE Robotics & Automation Magazine, 19(2), 98-100, 2012.

End-to-End Contextual Speech Recognition With Word-Piece-Level Token Selection

Zhibin Wu, Yang Zou*, Jian Zhou, Min Wang, Xiaoqin Zeng
Institute of Intelligence Science and Technology, School of Computer and Information,
Hohai University, Nanjing, China
{211307040022, yzou, 211607010098, mwang, xzeng}@hhu.edu.cn

***Abstract*—The utilization of dynamic contextual information in end-to-end automatic speech recognition has been an active research topic. Generally, the popular Contextual LAS (CLAS) provides favorable all-neural solutions. Nevertheless, it cannot be extended to large bias lists without many cases of recognition errors caused by similar pronunciation or word fragment repetition. To address this limitation, this paper proposes a model called Fine-CLAS on the basis of CLAS, which exploits word-piece-level contextual knowledge and fuse it with the original phrase-level contextual knowledge to enable the contextual bias module to focus on fine-grained contextual information. First, the prefix tree constraint is presented to reduce the number of contextual phrases. Then, a strategy for word-piece-level token selection is designed to obtain the new word-piece-level embedding vector. Finally, a contextual transformation chain is constructed between the word-piece-level embedding vector key-value pairs to attain new key-value pairs. The proposed model with these techniques can reduce the word error rate (WER) by 5.37% and 2.10%, and the F1-score by 1.10% and 2.10% on the datasets test-clean and test-other of LibriSpeech, demonstrating preferable ASR and contextual bias performance.**

Keywords-dynamic contextual information; end-to-end; all-neural; word-piece-level contextual knowledge

I. INTRODUCTION

We can always feel the convenience of speech recognition technology in our lives, such as in the most commonly used smartphones, smart appliances, wearable devices, voice navigation and in-car systems [1]. In such applications, speech recognition performance can be significantly improved by incorporating information about the speaker's context into the recognition process [2]. Examples of contextual information include the status of the conversation (e.g. words such as "stop", "cancel", etc.), the location of the speaker (e.g. "restaurant", "airport", etc.) [3], personalized information about the user (e.g. contacts, song playlists, etc.) [4], and other specific nouns.

In recent years, many end-to-end automatic speech recognition (ASR) methods, such as Connectionist Temporal Classification (CTC) [5,6], Recurrent Neural Network Transducer (RNN-T) [7-12], and Attention-based Encoder-Decoder (AED) [13-18], have been widely used in life. However, the recognition of context-specific phrases in these scenarios still needs to be improved as most contextual content is scarce in the training data.

In the current work, we still consider techniques that dynamically incorporate contextual information into the recognition process. In end-to-end systems, an approach can be implemented by performing log-linear interpolation between the E2E model and the n-gram language model (LM) at each step of the beam search [14, 19-24], without adding any other neural network, which is referred to as Shallow Fusion according to the terminology in [25]. However, re-scoring using an externally trained language model independently runs counter to the benefits obtained from the joint optimization of components from sequence-to-sequence models. Thus, Golan Pundak et al. [26] proposed Contextual-LAS (CLAS), a novel all-neural mechanism that exploits contextual information (provided as a list of contextual phrases) to improve recognition performance. The technique first embeds each contextual phrase (via tokenizers, sliced into a series of word piece units) into a fixed dimensional representation, and then uses an attention mechanism to focus on the available context during decoding. In addition, a number of contextual phrases are allowed during inference. Although the full neural context approach outperforms shallow fusion, it still suffers from a problem: the performance of the model drops significantly when dealing with hundreds or even thousands of contextual phrases, which is caused by the large number of contextual phrases with similar pronunciation or partial word repetition.

To solve the problem, improvements to the CLAS model are necessary. Sun et al. [27] proposed a Tree Constrained Pointer Generator (TCPGen) component that makes full use of prefix tree selection to narrow down candidate words, enables token units at the word-piece-level, and models attention on word piece. Following this line of thought, an observation can be made that incorporating word-piece-level contextual information into the CLAS model might be a feasible way to alleviate the problems caused by word fragment repetition.

Based on this observation, this paper proposes three techniques to improve the CLAS model: prefix tree constraint, word-piece-level token selection, and contextual transformation chain construction, and the improved model is referred to as Fine-CLAS. Unlike the previous CLAS model [26] which only contextually modelled phrase-level embeddings, we propose to fuse contextual information at two different levels, phrase-level

*Corresponding author: yzou@hhu.edu.cn (Y. Zou)

DOI reference number: 10.18293/DMSVIVA23-003

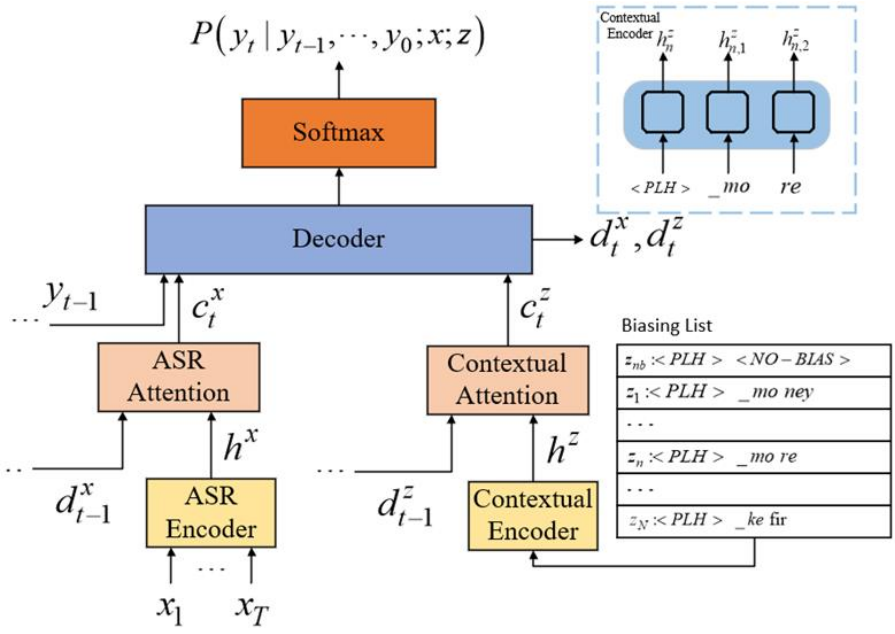


Figure 1. CLAS model: 1) The left-hand structure is the ASR of LAS and the right-hand is the context processing network; 2) The upper right-hand corner shows how the context encoder encodes a phrase and outputs its phrase embedding h_n^z and all the token embeddings $\left[h_{n,1}^z, h_{n,2}^z\right]$

embeddings and word-piece-level embeddings, in order to enable the contextual bias module to focus on fine-grained contextual information to match the ASR word-piece-level token output distribution.

The technical contributions of this paper are summarized as follows:

First, a prefix tree is constructed and combined with historical information to select whether to enable each phrase in the context list, which can reduce the number of phrases and obtain a smaller number of phrase-level biased embeddings and word-piece-level biased embeddings.

Second, a word-piece-level token selection algorithm is designed to select top-K phrases based on the weights and obtain the corresponding word-piece-level bias embeddings, which can result in a series of word-piece-level embedding information.

Third, a transformation chain between word-piece-level bias embeddings is constructed so as to obtain the transfer relationship between word-piece-level bias embeddings.

Fourth, compared to CLAS, the Fine-CLAS model constructed by incorporating the proposed techniques reduces word error rates (WER) by 5.37% and 2.10% and F1-scores (F1) by 1.10% and 2.10% on the test-clean and test-other test sets of LibriSpeech, where the list of contextual phrases consists of rare long-tail words. Furthermore, the Fine-CLAS model remains lightweight and modular, allowing for quick modifications to the contextual bias module without retraining the ASR model.

The rest of the paper is organized as follows: In Section 2 the standard AED model and the CLAS model are reviewed. In Section 3 the three techniques for improvement are described in detail. In Section 4 the experiment is described, followed by a discussion of the experimental results in Section 5. Finally, conclusions are presented in section 6.

II. BACKGROUND

A. *Attention-based Encoder-Decoder*

A standard AED contains three components: an encoder, a decoder and an attention network, as shown in the left-hand structure of Fig. 1. The encoder encodes the input $x_{1:T}$ as a sequence of high-level features h^x. In each decoding step t, the attention mechanism is utilized to combine the encoder output sequence into a single context vector c_t^x, which is used as part of the decoder input. The decoder is computed as follows.

$$d_t^x = Decoder\left(y_{t-1}, d_{t-1}^x, c_t^x\right) \quad (1)$$

where $Decoder(\bullet)$ denotes the decoder network and y_{t-1} is the embedding of the previous subword unit. The posterior distribution can be estimated using the Softmax output layer.

$$P\left(y_t \mid y_{t-1}, \cdots, y_0; x_{1:T}\right) = Soft\max\left(W^o\left[d_t^x; c_t^x\right]\right) \quad (2)$$

where $[\bullet;\bullet]$ denotes the splicing of two vectors. In the inference stage, the recognition result $y_{1:N}^*$ is calculated by performing beam search. In addition, shallow fusion [14,19-25] can be achieved by log-linear combination, as shown in the following equations.

$$y_{1:N}^* = \arg\max_{y_{1:N}} \log P\left(\mathrm{y}_{1:N} \mid x_{1:T}\right) + \lambda \log P^{LM}\left(y_{1:N}\right) \quad (3)$$

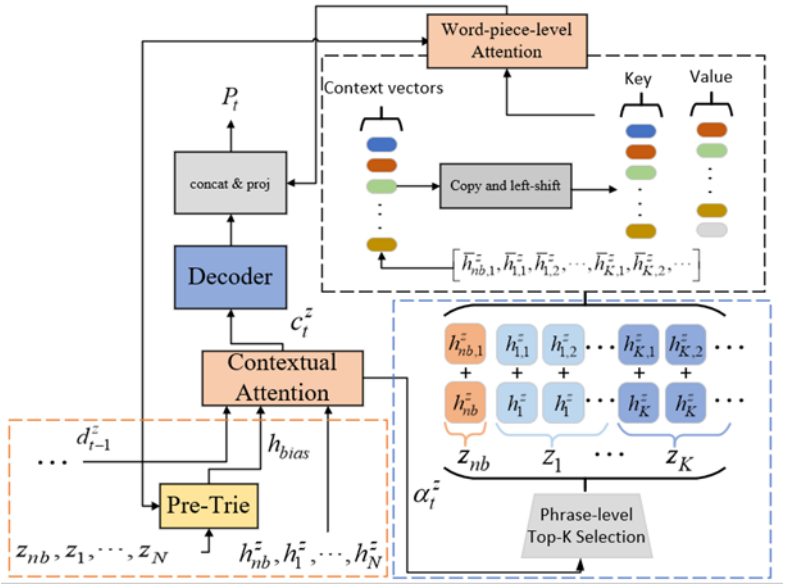


Figure 2. The structure of the Fine-CLAS model. Based on CLAS, it includes three additional modules: the prefix tree constraint, the word-piece-level token selection, and the contextual transformation chain construction, which are enclosed in the orange, blue, and black dashed boxes, respectively.

where λ is the hyperparameter controlling the relative importance of the LM output probability $P^{LM}(y_{1:N})$.

B. CLAS

CLAS models attention to contextual information, as shown in Fig. 1. The bias encoder embeds a list of biased phrases $Z=\{z_{nb}, z_1, z_2, \cdots, z_N\}$ into a set of vectors $h^z=\{h_{nb}^z, h_1^z, \cdots, h_i^z, \cdots, h_N^z\}$, where h_i^z is an embedding of z_i and $<PLH>$ is a phrase-level placeholder that represents the entire contextual phrase. Since biased phrases may be irrelevant to the current discourse, we introduce the phrase-level unbiased option z_{nb}. The embedding h_i^z is created by feeding a sequence of subword embeddings in z_i (i.e. the same lexical elements or chunk units used by the decoder) to the biased encoder and representing the whole phrase using the first state output of the LSTM. Attention modelling is then performed at h^z, using the decoder state d^t to compute the auxiliary context vector c_t^z. This context vector summarizes z at time step t and is calculated as shown below.

$$u_{it}^z = v^{z^T} \tanh\left(W_h^z h_i^z + W_d^z d_t + b_a^z\right) \quad (4)$$

$$a_t^z = soft\max\left(u_t^z\right) \quad (5)$$

$$c_t^z = \sum_{i=0}^{N} a_{it}^z h_i^z \quad (6)$$

Next, the context vector c_t^x, obtained by combining the ASR attention, yields the LAS context vector $c_t = \left[c_t^x; c_t^z\right]$ for the input decoder. It is worth noting that, given the audio and the previous output, CLAS can obtain the weights of the bias phrases that are of interest during the current decoding process, as follows.

$$a_t^z = P(z_t \mid d_t) = P(z_t \mid x; y_{<t}) \quad (7)$$

We refer to a_t^z as bias-attention-probability.

III. METHODS

The Fine-CLAS model is established on the CLAS model by augmenting three additional models that correspond to three approaches, as shown in Fig. 2. First, the prefix tree constraint is introduced to reduce the number of contextual phrases. Then word-piece-level token selection is performed to obtain the new word-piece-level embedding vector. Finally, a contextual transformation chain construction is executed between the word-piece-level embedding vector key-value pairs (K and V) to obtain new key-value pairs, which are used in the computation of the word-piece-level attention mechanism to obtain the final word-piece-level context vector.

A. Prefix Tree Constraint

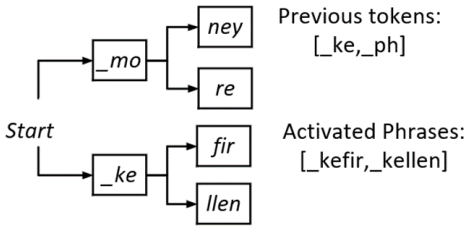


Figure 3. An example of prefix tree search.

In this subsection, we propose a trie-based bias module that encodes the bias list into a prefix tree at the word-piece-level, as shown in Fig. 3. Given the previously output word fragment tokens as queries, a certain history interval is selected and input to the bias module to find the phrases corresponding to the prefixes, returning a binary vector $h_{bias}=[a_0,a_1,\cdots,a_N]\in\{0,1\}$, with N being the number of phrases. $a_n=0$ means the phrase is not activated and not relevant to the sentence; $a_n=1$ means the phrase is activated and relevant to the sentence. h_{bias} is computed to filter relevant phrases and will only be used for phrase-level attention in the inference stage, as shown in the orange dashed box in Fig. 2.

B. Word-piece-level Token Selection

In this subsection, we propose a word-piece-level token selection technique. It introduces word-piece-level context vectors that are spliced and mapped to the decoder's output, thus matching the token units of ASR with word fragments as the output distribution and reducing the uncertainty of token prediction, as shown in the blue dashed box in Fig. 2.

First, the token-level acoustic embedding vector d_{t-1}^z for the current time step t is modeled with a series of phrase-level bias embedding vectors $h^z=\left[h_{nb}^z,h_1^z,\cdots,h_N^z\right]$ for phrase-level attention, resulting in phrase-level context weights $a_t^z=\left[a_{t,nb}^z,a_{t,1}^z,\cdots,a_{t,N}^z\right]$. Then the average attention weight $\breve{a}_t^z=\left[\breve{a}_{t,nb}^z,\breve{a}_{t,1}^z,\cdots,\breve{a}_{t,N}^z\right]$ is calculated based on the global (time step t with all previous attention) or local (time step t with the attention of the previous finite time step). The size of the list of context-biased phrases can be hundreds or thousands, which is not small even after prefix tree filtering. If we directly use the word-piece-level embedding vector for each phrase, the corresponding list of word-piece-level embedding vectors will become very large. So we select top-K attention weights from $\breve{a}_t^z$, and then get the corresponding contextual bias phrases according to the index of the selected weights to achieve the reduction from N to K. For each bias phrase selected, the first state output h_k^z of the encoder representing the phrase-level embedding vector is respectively added to the subsequent state output $h_{k,i}^z$ of the encoder representing the word-piece-level embedding vector, and we get a series of word-piece-level embedding vectors corresponding $\bar{h}_{k,i}^z$, which results in a list of all word-piece-level embedding vectors $K=V=\left[\bar{h}_{nb,1}^z,\bar{h}_{1,1}^z,\bar{h}_{1,2}^z\cdots,\bar{h}_{K,1}^z,\bar{h}_{K,2}^z,\cdots\right]$. The specific formula is as follows.

$$\left[z_{nb},z_1,\cdots,z_K\right]=PhraseTopKSelection\left(Z,\left[\breve{a}_{t,nb}^z,\breve{a}_{t,1}^z,\cdots,\breve{a}_{t,N}^z\right]\right) \quad (8)$$

$$\left[h_k^z,h_{k,1}^z,h_{k,2}^z,\cdots\right]=ContextualEnc\left(z_k\right) \quad (9)$$

$$\left[h_{nb}^z,h_{nb,1}^z\right]=ContextualEnc\left(z_{nb}\right) \quad (10)$$

$$\bar{h}_{k,i}^z=h_k^z+h_{k,i}^z \quad (11)$$

$$\bar{h}_{nb,1}^z=h_{nb}^z+h_{nb,1}^z \quad (12)$$

$$K=V=[\bar{h}_{nb,1}^z,\bar{h}_{1,1}^z,\bar{h}_{1,2}^z,\cdots,\bar{h}_{K,1}^z,\bar{h}_{K,2}^z,\cdots] \quad (13)$$

C. Contextual Transformation Chain Construction

Although in the word-piece-level token selection technique, word-piece-level contexts are constructed for use in the decoding step to achieve fine-grained local bias, the probability of transfer between word fragment tokens are not explicitly modelled. Modelling this transfer may be helpful when the context is personalized entity names and proper names that are rare or invisible during training, as it allows us to recover the expected next token by using the preceding subsequence. We therefore introduce a more fine-grained biasing technique that operates at the word-piece-level, following word-piece-level token selection, as shown in the black dashed box in Fig. 2.

Specifically, we construct an associative memory to store and retrieve the associated bias context. As shown in Fig. 2, the memory stores association transfers between word-piece-level subsequences of the same phrase. In the associative memory, the key of each word-piece-level token in each phrase is mapped to the value of the next word-piece-level token (left shift). The original formula for the key-value pair selected by the word-piece-level token is as follows.

$$k_l=v_l=\bar{h}_{k,i}^z \quad (14)$$

Accordingly, the memory entries of the key-value pair (k_l,v_l) constructed after the contextual transformation chain are two consecutive word-piece-level embedding vectors $\bar{h}_{k,i}^z$ and $\bar{h}_{k,i+1}^z$, as follows.

$$(k_l,v_l)=\left(\bar{h}_{k,i}^z,\bar{h}_{k,i+1}^z\right) \quad (15)$$

IV. Experiments

A. Datasets and Metrics

Our experiments are conducted on the dataset Librispeech. The dataset is collected from an audiobook website, and speech recognition is done once for each sentence. The acoustic model from the WSJ example is adopted as the recognition model, a binary grammar is utilized as the language, and the input dataset for the language model is the e-book text corresponding to the speech data. From the clean data, 20 males and 20 females are randomly selected as the development set (dev-clean), the remaining speakers are selected as a test set of the same size (test- clean), and the rest as the training set. The training set is 100 hours (train-clean-100). In the other data, the WERs are sorted from lowest to highest, and the test set is randomly selected near the third quartile (test-other). As LibriSpeech's test set lacks a bias list, we construct a bias list by collecting words other than the 20,000 most common words in the training data from the reference of the test set and discarding short words of

less than 5 letters. Finally, the simulated bias lists for test-clean and test-other consists of around 1,000 phrases.

Firstly, a set of evaluation metrics is introduced that tracks three different aspects of ASR, (1) WER: overall word error rate assessed for all words, (2) CER: overall character error rate assessed for all words, (3) U-WER: unbiased word error rate assessed for words not in the bias list. Secondly, contextual bias is measured using the precision (P), recall (R) and F1-score (F1) of the biased phrases. In summary, we use six evaluation metrics to measure the performance of Fine-CLAS.

B. Configurations

The model evaluated in this paper is trained on an A40 graphics card with 48G of video memory and a batch size of 8. To improve the performance of the model, the data enhancement method of SpecAugment is used. The input features are a 40-dimensional log-mel filter bank with a sampling rate of 16000Hz, extracted from a window of length 25ms, length of the hop of the sliding window is 10ms, and its output vocabulary is a 1000-word block generated via BPE.

The ASR encoder is composed of a convolutional module, a cyclic module and a fully connected module. The convolutional module consists of two 3x3 convolutional layers with 128 and 256 nodes, the cyclic module includes four bi-directional LSTM layers with 1024 nodes each and the fully connected module comprises two fully connected layers with 512 nodes. The ASR encoder attention is computed in 1024 dimensions using a content-based attention mechanism. The decoder contains 1 GRU with 1024 nodes. The context encoder involves 1 bi-directional LSTM layer with 128 nodes, and the phrase-level attention and word-piece-level attention have the same structure as the ASR attention.

The model has a total of 177.4M trainable parameters and our model is implemented using Pytorch and Speechbrain.

In order to exercise the "no bias" option, we use the same settings as in [26]. In all experiments, we set $P_{keep} = 0.5$ to improve robustness to the "no bias" case, and set $N_{phrase} = 1$ and $N_{order} = 4$. This results in an expected size of 5 for the bias list (half the batch size, plus one "no bias" option). In addition, the phrase selection has K of 5. In inference, we adopt a beam size of 10 for the search.

V. RESULTS

A. Evaluation Results for ASR

TABLE I. ASR test results on test-clean and test-other

Model	test-clean			test-other		
	WER	CER	U-WER	WER	CER	U-WER
AED	21.79	10.98	19.90	45.65	26.35	41.30
CLAS	22.97	16.37	19.90	38.18	22.63	33.90
Fine-CLAS	**17.60**	**10.63**	**16.00**	**36.08**	**20.95**	**32.70**

We use the simulated bias list to validate the improvements to the model, and evaluate the performance of the model ASR on three metrics. As shown in Table I, AED achieves a word error rate of 21.79% on test-clean and 45.65% on test-other, which is the result of testing without additional language model. Compared with AED, the improved CLAS model shows an increase in WER and CER on test-clean and a noticeable decrease in WER and CER on test-other, indicating that the ASR performance of the CLAS model is quite good. Compared with CLAS, our improved Fine-CLAS model decreases WER by another 5.37% and 2.10% on test-clean and test-other, respectively, and achieves noticeable improvements in the other two metrics. This indicates that the ASR performance of our model is preferable.

B. Evaluation Results for Contextual Biasing

TABLE II. Contextual bias test results on test-clean and test-other

Model	test-clean			test-other		
	F	R	F1	F	R	F1
AED	**97.10**	37.80	54.40	82.40	15.60	26.20
CLAS	92.90	59.80	72.80	88.80	29.40	44.10
Fine-CLAS	96.00	**66.10**	**73.90**	**95.50**	**30.40**	**46.20**

To test the effectiveness of the proposed model's contextual bias, we use three evaluation metrics, as shown in Table Ⅱ. First, the CLAS model achieves better performance than the AED, especially in the F1-score metric, which is improved by almost 20%, indicating that the CLAS model noticeably improves the contextual bias effect. Compared to CLAS, our Fine-CLAS model achieves a slight improvement with another 1.10% and 2.10% improvement in F1-score on test-clean and test-other, respectively. This indicates that our model improves both the performance of the ASR model and the effect of contextual bias.

VI. CONCLUSION

In this work, we propose the Fine-CLAS model that promotes end-to-end contextual speech recognition through three techniques: prefix tree constraint, word-piece-level token selection, and contextual transformation chain construction. The improved model can mitigate confusion caused by similar pronunciations or word fragment repetition. The experimental results of several evaluation metrics on the dataset LibriSpeech clearly show that these proposed techniques improve the performance of the original context-biased approach and make the Fine-CLAS model more capable of handling a large number of contextual phrases. In the future work, we shall attempt to further expand the context bias list and explore even better methods for dealing with contextual issues. In addition, we shall attempt to combine it with ChatGPT, an AI chatbot, to explore multimodal contextualization from speech to text.

REFERENCES

[1] I. McGraw et al., "Personalized speech recognition on mobile devices," IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5955-5959, 2016.

[2] P. Aleksic, M. Ghodsi, A. Michaely, et al., "Bringing Contextual Information to Google Speech Recognition," Proc. Interspeech, pp. 468-472, 2015.

[3] J. Scheiner, I. Williams and P. Aleksic, “Voice search language model adaptation using contextual information,” IEEE Spoken Language Technology Workshop, pp. 253-257, 2016.

[4] P. Aleksic, C. Allauzen, D. Elson, A. Kracun, D. M. Casado and P. J. Moreno, “Improved recognition of contact names in voice commands,” IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5172-5175, 2015.

[5] A. Graves, S. Fernández, F. Gomez, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” Association for Computing Machinery, pp. 369-376, 2006.

[6] A. Graves, N. Jaitly, “Towards end-to-end speech recognition with recurrent neural networks,” International Conference on Machine Learning, pp. 1764-1772, 2014.

[7] A. Graves, “Sequence Transduction with Recurrent Neural Networks,” Computer Science, arXiv preprint arXiv:1211.3711, 2012.

[8] A. Graves, A.R. Mohamed, G. Hinton, “Speech recognition with deep recurrent neural networks,” IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6645-6649, 2013.

[9] R. Prabhavalkar, K. Rao, T. Sainath, B. Li, L. Johnson & N. Jaitly, “A comparison of sequence-to-sequence models for speech recognition,” Proc. Interspeech, pp. 939-943, 2017.

[10] E. Battenberg, J. Chen, R. Child, et al., “Exploring neural transducers for end-to-end speech recognition,” IEEE Automatic Speech Recognition and Understanding, pp. 206-213, 2017.

[11] J. Li, R. Zhao, H. Hu and Y. Gong, “Improving RNN Transducer Modeling for End-to-End Speech Recognition,” IEEE Automatic Speech Recognition and Understanding, pp. 114-121, 2019.

[12] Y. He et al., “Streaming End-to-end Speech Recognition for Mobile Devices,” IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6381-6385, 2019.

[13] J. Chorowski, D. Bahdanau, D. Serdyuk, et al., “Attention-Based Models for Speech Recognition, ” Neural Information Processing Systems, pp. 577-585, 2015.

[14] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel and Y. Bengio, “End-to-end attention-based large vocabulary speech recognition,” IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4945-4949, 2016.

[15] L. Dong, S. Xu and B. Xu, “Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition,” IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5884-5888, 2018.

[16] C.C. Chiu, T.N. Sainath, Y. Wu, et al., “State-of-the-art Speech Recognition With Sequence-to-Sequence Models,” IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4774-4778, 2018.

[17] A. Zeyer, K. Irie, R. Schlüter, et al., “Improved training of end-to-end attention models for speech recognition,” Proc. Interspeech, pp. 7-11, 2018.

[18] W. Chan, N. Jaitly, Q. Le and O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition, ” IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4960-4964, 2016.

[19] I. Williams, A. Kannan, P. Aleksic, D. Rybach, T. Sainath, “Contextual speech recognition in end-to-end neural network systems using beam search,” Proc. Interspeech, pp. 2227-2231, 2018.

[20] Z. Chen, M. Jain, Y. Wang, M. L. Seltzer and C. Fuegen, “End-to-end Contextual Speech Recognition Using Class Language Models and a Token Passing Decoder,” IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6186-6190, 2019.

[21] D. Zhao, T. Sainath, D. Rybach, P. Rondon, D. Bhatia, B. Li & R. Pang, “Shallow-fusion end-to-end contextual biasing,” Proc. Interspeech, pp. 1418-1422, 2019.

[22] R. Huang, O. Abdel-Hamid, X. Li, et al., “Class LM and word mapping for contextual biasing in End-to-End ASR,” Proc. Interspeech, pp. 4348-4351, 2020.

[23] Y.M. Kang, Y. Zhou, “Fast and Robust Unsupervised Contextual Biasing for Speech Recognition,” arXiv preprint arXiv:2005.01677, 2020.

[24] C. Liu, D.R. Liu, F. Zhang, et al., “Contextualizing ASR Lattice Rescoring with Hybrid Pointer Network Language Model,” Proc. Interspeech, pp. 3650-3654, 2020.

[25] C. Gulcehre, O. Firat, K. Xu, et al., “On Using Monolingual Corpora in Neural Machine Translation,” arXiv preprint arXiv:1503.03535, 2015.

[26] G. Pundak, T. N. Sainath, R. Prabhavalkar, A. Kannan and D. Zhao, “Deep Context: End-to-end Contextual Speech Recognition,” IEEE Spoken Language Technology Workshop, pp. 418-425, 2018.

[27] G. Sun, C. Zhang and P. C. Woodland, “Tree-Constrained Pointer Generator for End-to-End Contextual Speech Recognition,” IEEE Automatic Speech Recognition and Understanding Workshop, pp. 780-787, 2021.

A BERT-based Model for Semantic Consistency Checking of Automation Rules

Bernardo Breve, Gaetano Cimino, Vincenzo Deufemia, Annunziata Elefante
Department of Computer Science
University of Salerno, Italy
{bbreve, gcimino, deufemia, anelefante}@unisa.it

Abstract

Trigger-Action Platforms (TAPs) allow users to automate behaviors involving IoT devices either by programming rules from scratch or by accessing a catalog of user-defined rules. Users can search the catalog based on their interests and needs, browsing through rules that are expressed according to textual descriptions supplied by the rule's creator. However, TAPs do not perform any control over these User-defined descriptions (UDDs), which means that there is no way to ensure their suitability. This lack of control might lead to the inclusion of erroneous information, making it challenging for users to retrieve the relevant rules they need during searches. To address this issue, this paper proposes the use of a BERT-based classification model to check the semantic consistency of a rule's UDD with respect to its trigger-action components. We evaluate the proposed solution with the popular TAP, namely If-This-Then-That (IFTTT), by training the model on a dataset consisting of 9643 labeled samples. Each sample is composed of a pattern derived from the rule components, the corresponding rule's UDD, and a label expressing whether they are semantically related. The code of the software is publicly available on GitHub[1].

Index terms— Trigger-action rules, Semantic consistency checking, NLP, BERT, IoT platforms.

1 Introduction

The outburst of Internet-of-Things (IoT) leads to a new world of opportunities and challenges for programmers and end-users who use this technology [15]. In fact, these "smart" devices are spreading across houses in the form of sensors and actuators, such as cameras, lights, thermostats, locks, and so forth. Thus, an IoT device can sense the environment around it, acquiring data and sharing them over the internet with other devices, leading to an ecosystem of devices that can collaborate to generate automation [2].

To empower users in getting the most out of IoT devices by defining useful tasks such as lights turning off automatically at sunset, IoT-based applications are built, which help users define interoperability behaviors in a simple way [11]. Among the different types of platforms, the most popular are the *Trigger-Action Platforms* (TAPs), which empower users to define custom behaviors by means of conditional rules [9], consisting of a trigger component, which specifies the event whose occurrence would trigger the rule, and an action component, which describes the operation to perform in order to accomplish the behavior. In addition, most platforms allow users to specify additional information in the form of textual description, called *User-defined description* (UDD), which summarizes the behavior of the rule.

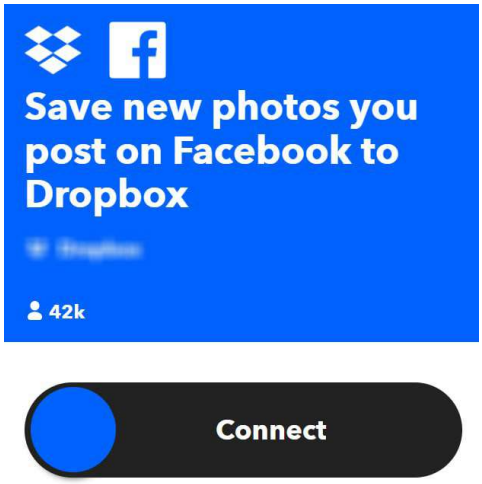


Figure 1: An applet's example with the associated UDD

If-This-Then-That (IFTTT)[2] is the most popular TAP. It was first released in 2010 and since then, it has gained a

[1] https://github.com/empathy-ws/Semantic-Consistency-Checking-of-Automation-Rules

DOI reference number: 10.18293/DMSVIVA23-008

[2] https://ifttt.com

larger number of followers. One advantage of IFTTT is the vast catalog of rules (known as *applets*) that users can access, which have been created and shared by other members of the community. In this context, UDDs become even more critical, as they help users to easily understand the behavior of applets while browsing the catalog. Figure 1 showcases an example of an IFTTT applet from the applet catalog (with the author's name blurred for privacy), which is presented based on its UDD. The UDD summarizes the applet's behavior, which in this case consists of automatically synchronizing any new photo posted by the user on Facebook to a Dropbox folder.

Unfortunately, the literature has highlighted that IoT smart environments are not risk-free with respect to security and privacy concerns [21], as devices might represent a suitable target for malicious individuals to plan cyberattacks [1]. Furthermore, users themselves can induce cyber security threats while interacting with TAPs [17]. In fact, rules created through TAPs possess inherent risks, mainly caused by the level of technical knowledge that the average TAP user has, which might not be sufficient enough to understand the severeness of an apparently innocuous rule [6]. For example, a rule such as "If I enter the gym, post a tweet with my account", might at first glance appear as an innocuous rule, however, providing a malicious individual with a routine produced by the rule, might give him/her useful insights into when planning a possible theft, aware that the house of the user will be empty. To address these concerns, the research has moved towards the definition of ad-hoc solutions for empowering users in protecting the smart environment and their privacy [3, 4, 5, 18].

The existence of fields such as UDDs also represents an important concern for users, which still has not received much attention from the literature. In fact, TAPs such as IFTTT do not perform actual control over what the authors of rules may write as a UDD, leaving them the freedom to type anything they want for describing the behavior. This may represent an issue for several reasons. First, a rule's creator may type a UDD completely unrelated not only to the behavior of the rule but also to the characteristics that a description typically should have, e.g., "You will like this rule!", making it virtually impossible for a user to find such a rule. Also, a rule with an imprecise UDD might force it to appear in the search results for other types of rules, making it even more complicated for users to identify rules that meet their needs. Finally, rules might not be understood when provided with poor UDDs, as the latter represents their showcase.

Therefore, in this paper, we address the above-mentioned concerns by proposing a classification model based on *Bidirectional Encoder Representations from Transformers* (BERT) [8] to determine the degree of semantic consistency established between the behavior of a rule and its associated UDD. To perform this comparison, we devised a methodology where the actual behavior is encapsulated within a textual pattern constructed from the rule components. The resulting pattern, along with the rule's UDD, is given as input to the classification model, which computes a semantic similarity score between the two texts. The proposed solution was evaluated on rules gathered from the IFTTT platform, using a dataset of 9643 UDD-pattern labeled pairs. The empirical analysis shows that the model effectively categorized semantic consistency, achieving an accuracy rate of approximately 92%.

The paper is organized as follows: Section 2 discusses the state of the art on semantic analysis of automation rules. Section 3 outlines the overall methodology by focusing on the dataset, the labeling process, and the model architecture. Then, in Section 4, we describe the experimental evaluation leading to the performance scores we traced back from the model. Finally, Section 5 concludes the manuscript and provides future directions for our proposal.

2 Related work

In this section, we describe the primary research efforts related to the semantic analysis of IF-THEN rules. Previous studies in the literature have focused on language-to-code methodologies, which extract executable code from rule descriptions. On the other hand, alternative studies aim to enhance the user experience by developing sophisticated graphical interfaces or by utilizing sequence-to-sequence models to automatically generate rule components, thereby simplifying the rule creation process for users.

The use of natural language to program computers could potentially increase accessibility to modern technology for inexperienced users [13]. In this regard, developing language-to-code translators could help create IF-THEN rules that cater to user needs. This may be accomplished by leveraging a semantic parser that converts natural language into executable code, thereby streamlining the process of applet customization and making it more user-friendly for a broader range of users. In [16], Quirk *et al.* designed a language-to-code approach for natural language programming. Specifically, the authors collected 114,408 applet-description pairs from the IFTTT website and used them to train semantic parser learners that could effectively interpret natural language descriptions of applet behaviors and map them to executable code. The IF-THEN statements were represented as syntactic constructs through the use of Abstract Syntax Trees (ASTs), where each node denoted a specific text construct and captured its structural and content-related details. The constructed ASTs were then fed to several classifiers, which iteratively searched for the most likely derivation, adding correct instances to the set of positive instances, and incorrect instances to

the set of negative instances. The classifiers were then retrained using the revised training data, and the process was repeated until the desired performances were achieved or a maximum number of iterations was reached. In [12], Chen *et al.* presented a neural network architecture for automatically translating natural language descriptions to IF-THEN rules. Specifically, the authors designed an attention architecture, called *Latent Attention*, that computes the importance of each word in the description for predicting rule components in a two-stage process. Yusuf *et al.* proposed *RecipeGen*, a deep learning-based approach that uses a Transformer sequence-to-sequence architecture to generate IF-THEN rules from natural language descriptions [20]. The problem is modeled as a sequence learning and generation task, which facilitates the abstraction of implicit relations between rule components. To improve the generation performance, RecipeGen relies on autoencoding pre-trained models to initialize the parameters of the encoder in the sequence-to-sequence model.

It is noteworthy that previous studies have limited their scope to interactions requiring a user's request and the system's response in the form of an interpretation. An essential aspect is to involve the user in an interactive dialogue to validate and improve their intention and create a complete and accurate rule. Concerning this matter, in [7], Corno *et al.* proposed *HeyTAP*, a conversational and semantic-powered platform that can map abstract user needs to executable IF-THEN rules. HeyTAP uses a multimodal interface to interact with the user and extract personalization intentions for different contexts. The authors conducted an exploratory experiment with 8 users to test the effectiveness of HeyTAP in guiding participants from abstract needs to actual IF-THEN rules. Results showed that HeyTAP can successfully translate abstract user needs into IF-THEN rules that can be executed by contemporary TAPs. While previous semantic parsers, as presented, perform both text parsing and comprehension in a single step, Yao *et al.* proposed an approach relying on a Hierarchical Reinforcement Learning framework to translate natural language descriptions into IFTTT applets [19]. This approach introduces an interactive element to semantic analysis, where an agent is trained with a hierarchical policy to maximize the parsing accuracy while minimizing the number of questions asked to the user. Finally, Huang *et al.* conducted a thorough analysis of the potential implications of incorporating natural language interfaces for assisting users in the customization and automation of their personal devices [10]. In particular, the authors introduced *InstructableCrowd*, a crowd-powered system that enables users to program their devices via a natural language interface. The system is based on two key design decisions: i) creating simple programs that are easy to use and ii) employing human crowd workers to operate the natural language interface instead of using automated systems. The system is oriented around relatively simple IF-THEN rules and contains more than one sensor/effector. The authors argue that InstructableCrowd addresses the main problems with device customization and automation, and could provide a new way to program devices in the future.

With respect to the approaches that focus on analyzing natural language descriptions to generate executable rules [7, 12, 16, 19, 20], or that investigate how to interact with users in order to improve the rule definition process [10], we address a different problem since we focus on checking the semantic consistency of a UDD against the actual rule behavior before its dissemination.

3 Methodology

In this section, we outline the methodology used to develop a model that evaluates the semantic consistency between the trigger-action components of an IFTTT applet and the natural language description provided by the creator. Specifically, we describe the dataset employed during the experimental evaluations and the technical details for implementing the semantic consistency evaluation model.

3.1 IFTTT Applet Dataset

In our study, we utilized the dataset proposed by *Mi et al.* [14], which consists of a collection of IFTTT applets obtained from crawling the IFTTT.com website. This dataset includes crucial information such as a title (*Title*), a description explaining the applet behavior (*Desc*), the event triggering the applet (*TriggerTitle*) defined through a specific channel (*TriggerChannelTitle*), the action to be performed (*ActionTitle*) selected from the corresponding channel (*ActionChannelTitle*), and the name of the applet creator (*Creator Name*). We employed the information generated by IFTTT to design a new pattern for *synthesizing* UDDs, ensuring a coherent and accurate representation. Then, we performed *dataset labeling*, a critical process that involved assigning suitable labels to the synthesized dataset to facilitate efficient categorization, organization, and analysis of the data.

3.1.1 Synthesizing a UDD from the components of an applet

To evaluate how consistent a UDD is with an applet's actual behavior, we designed a pattern that acts as a natural language description by leveraging applet key components, including trigger, trigger channel, action, and action channel. In particular, the pattern utilized for generating the synthesized UDD is as follows:

*IF TriggerTitle (TriggerChannelTitle) **THEN** ActionTitle (ActionChannelTitle)*

This standardized structure serves as a concise and comprehensive means of depicting the essential components and events associated with a specific applet. To further illustrate, we provide an example employing an IFTTT applet consisting of the following components:

- ***TriggerTitle***: "Any new SMS received"
- ***TriggerChannelTitle***: "Android SMS"
- ***ActionTitle***: "Send me an email"
- ***ActionChannelTitle***: "Email"

The pattern generated for this applet is as follows:

*IF Any new SMS received (Android SMS) **THEN** Send me an email (Email)*

This pattern provides a clear and concise representation of the applet's components and their corresponding values, enabling a comprehensive understanding of its intended behavior. This statement remains true even after reading the original description:

When a text message arrives, forwards it to your email.

3.1.2 Dataset labeling

Another crucial aspect is the construction of a set for the training phase of the proposed model. In this regard, we randomly selected a subset of applets and labeled them over a period of three weeks. The labels were selected based on the correlation between the UDD and the description synthesized by the pattern presented in the previous section. In particular, the UDD-pattern pairs were labeled according to the following similarity label values:

- ***contradiction***: denotes inconsistency between the UDD and the synthesized pattern.
- ***entailment***: denotes consistency between the UDD and the synthesized pattern.

The labeling of the pairs was determined based on the following conditions: If a UDD accurately depicted both the trigger and action components of a rule, it was assigned the label entailment. Otherwise, the UDD-pattern pair was labeled as contradiction. Specifically, we applied the majority method for manually labeling the pairs of the considered dataset. The first, second, and fourth authors were in charge of manually labeling the pairs, with the third author intervening in cases where there was no agreement. With this strategy, we obtained a dataset containing 10,543 labeled pairs, where 6,540 belong to class entailment and 4,003 to class contradiction. This careful labeling of pairs ensures that our model is trained on a diverse and representative set of data, enabling reliable evaluations and accurate assessments of semantic consistency between applet patterns and descriptions.

The resulting dataset was used to train and evaluate the BERT-based classification model adopted to perform the semantic consistency checking task.

3.2 The Proposed BERT-based Model

The architecture of the proposed model for classifying the semantic consistency of an applet's UDD with respect to the corresponding pattern is depicted in Figure 2. The model comprises several interconnected components working together to achieve our goal.

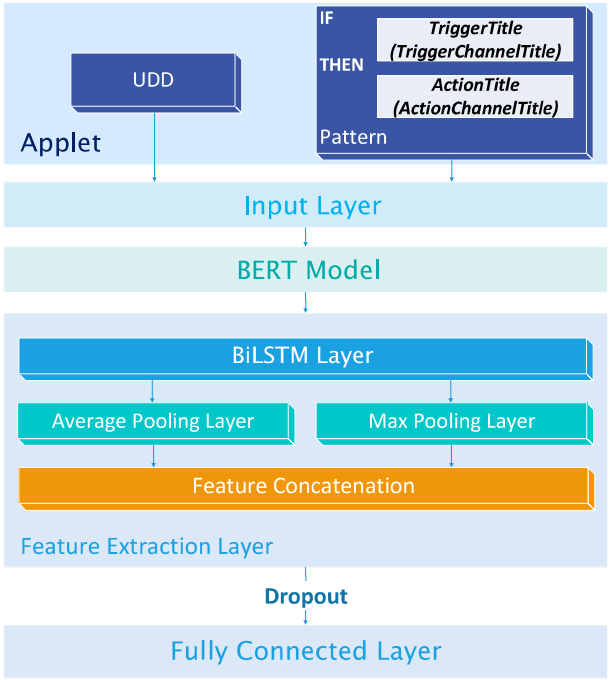


Figure 2: The architecture of the BERT-based model

The first component is an *Input Layer* that takes the UDD-pattern pairs from the dataset and encodes them into numerical representations, also known as dense-type vectors. Once the texts are transformed into dense vectors, they are passed to a state-of-the-art language model that has been pre-trained on a large corpus of text, namely BERT. The latter consists of multiple Transformer Encoder Layers that generate contextual representations of each word in

the input sequence by leveraging the self-attention mechanism. Each layer outputs a series of dense vectors that capture different levels of syntactic and semantic information. The next step involves passing the sequence obtained from the BERT model as input to a *Feature Extraction Layer*. This layer includes a *Bidirectional Long Short-Term Memory* (BiLSTM) *Layer*, which is a variant of the traditional LSTM Layer that is designed to store the both past and future context of a sequence. In this case, the BiLSTM Layer consists of 64 LSTM cells that are connected in a chain. The output of the BiLSTM Layer is a sequence of vectors, where each vector represents the hidden state of the LSTM cell at a given time step. These hidden states are then concatenated to form a representation of the input sequence that captures global features and dependencies. The output from the BiLSTM Layer has then proceeded through two pooling layers, i.e., an *Average Pooling Layer* and a *Max Pooling Layer*, which reduce the dimensionality of the input data by aggregating information across the sequence. In particular, the first type of pooling calculates the average value of each feature throughout the sequence, capturing the overall representation and distribution of the features. In this way, it can help mitigate the impact of outliers or extreme values in the sequence. On the other hand, the second one selects the maximum value from each dimension of the vectors, capturing salient and important feature values. Thus, it can help highlight the most relevant information and discard less important details, which can be beneficial for identifying key elements or detecting specific patterns. The resulting vectors are concatenated into a single one by a *Feature Concatenation* module, which produces a compact representation of the input texts. Specifically, average pooling provides a global representation of the sequence, while max pooling focuses on the most significant features. Therefore, concatenating the results of both pooling operations into one vector allows the model to have a comprehensive representation that captures both the overall context and important local details. Before feeding the concatenated data into the *Fully Connected Layer*, a *Dropout* operation is applied. This randomly drops out a fraction of the input features, preventing the model from relying too heavily on any single feature and mitigating the risk of overfitting. Finally, the Fully Connected Layer exploits the extracted features to evaluate the semantic consistency of the applet descriptions with respect to the corresponding patterns. In particular, it uses the vector obtained from the previous layer as input and applies a series of linear transformations to compute the final classification output of the model.

4 Experimental Evaluation

In this section, we present an analysis of the performances of the implemented model. Specifically, we provide details on the experimental setup, the adopted metrics, and the results obtained from the experiments.

4.1 Evaluation Setup

We trained the BERT-based model through a two-step process. Initially, we froze all the pre-trained layers and performed a training process solely targeting the top layers of the model. This enabled feature extraction by exploiting the representations of the pre-trained model. After the feature extraction process, we performed an additional fine-tuning step. This involved unfreezing the BERT model and retraining the entire architecture using a considerably low learning rate. The purpose of this step was to progressively adapt the pre-trained features to the new data, significantly enhancing the performances of the model.

To pre-train and tune the proposed model for our semantic consistency checking task, we used the Python libraries `Keras` and `TensorFlow`. Among the range of pre-trained BERT models currently available, we used the "bert-base-uncased" variant in our study. This model corresponds to the "base" version, featuring 12 transformer blocks, 768 hidden units, and 12 self-attention heads. Additionally, it is designed to account for lowercase letters. The training set included 6006 entailment pairs and 3637 contradiction pairs, totaling 9643 samples. Furthermore, in order to determine the best hyperparameter configuration, we employed a validation set consisting of 150 entailment pairs and 150 contradiction pairs, yielding as best values: 4 epochs, 32 as batch size, 1e-5 as epsilon, and 70 as a maximum text length. Finally, we evaluated the model's performances using a test set comprising 384 entailment pairs and 216 contradiction pairs.

4.2 Evaluation Metrics

The evaluation of the proposed model's performances is based on multiple metrics, including Accuracy, Precision, Recall, and F1-score. Specifically, the assessment metrics are derived from the values of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Accordingly, the evaluation metrics are formulated as follows:

- **Accuracy** is a measure of the overall correctness of a model's predictions, expressed as the ratio of the number of correctly classified instances to the total number of instances evaluated: $\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$
- **Precision** is a measure of the proportion of true positive instances among all instances that the model identified as positive: $\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}$

- **Recall** is a measure of the proportion of true positive instances among all actual positive instances: $\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$

- **F1-score** is the harmonic mean of Precision and Recall: $\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}}$

4.3 Results and Discussion

Figure 3 shows the confusion matrix obtained from the classification results on the test set, whereas Table 1 reports the resulting values of Accuracy, Precision, Recall, F1-score, and the average of the per-class metrics.

Figure 3: Confusion Matrix of the BERT-based model

We can observe that the model exhibits satisfactory performances in discriminating between entailment and contradiction UDD-pattern pairs, as highlighted by the Accuracy value of 92%. Upon analyzing the values for individual classes, it is feasible to note that the model primarily identifies entailment pairs, achieving Recall and F1-score values of 97% and 94%, respectively. Nevertheless, it is worth noting that the model tends to produce the entailment class more frequently than the contradiction class, leading to a lower Recall value for the contradiction class (83%). In particular, it wrongly classifies a contradiction pair as entailment 36 times, influencing the Precision value of this class (91%). An in-depth manual analysis of classification results reveals that this scenario is due to the structure of UDDs, which despite providing high knowledge about applet behaviors, may not include enough detail to understand them completely. As an example, consider the following pair:

> **Pattern:** *If a new photo post by you with a hashtag on Facebook then upload the file from the URL on Google Drive*
> **UDD:** *Crowdsource wedding photos from Facebook with a hashtag*

This pair is labeled as contradiction because, while presenting some correct information about the applet's behavior, the description does not perfectly explain the action to be performed. In this case, the model misclassifies the pair as entailment, probably due to the fact that it focuses significant attention on the actual relevant part of the text. In general, we can argue that the model might misjudge a pair when a user customizes a description based on how s/he will use the applet rather than specifying the components of the applet. On the other hand, we can observe that the model shows high reliability when classifying a pair with the contradiction class, as proved by the few cases where an entailment pair is classified as contradiction, i.e., 13 times, resulting in a high Precision value for this class (93%).

Table 1: Classification performances on the test set

Metric	Contradiction	Entailment	Avg
Precision (%)	93	91	92
Recall (%)	83	97	90
F1-score (%)	88	94	91
Accuracy (%)			92

5 Conclusion and Future Work

TAPs have enabled all types of users to easily define complex automation related to IoT devices by means of simple conditional rules. Such rules can then be described through UDDs freely typed by users without there being an effective check by the platforms on the correctness of what is written. Thus, in this paper, we addressed this issue by proposing a BERT-based classification model for evaluating the semantic consistency between the UDD of a rule and its actual behavior, the latter inferred by analyzing the trigger and action rule components. Our experimental evaluation over a case study on the IFTTT platform, considering a dataset of 9643 manually labeled UDD-pattern pairs, revealed an overall accuracy rate of 92%, indicating high reliability of the model in discriminating a compliant UDD with respect to an unrelated one.

In the future, we would like to consider the adoption of Large Language Models (LLMs) to generate, from the rules components, ad-hoc descriptions which might be presented to the user as a suggestion, perhaps in place of a less

suitable UDD. Furthermore, we might consider introducing additional classification outputs for the model, indicating a description that, although it contains some correct information, it is not sufficiently detailed to make clear the behavior of the rule from the perspective of both the trigger and action components.

Acknowledgements

This work has been supported by the Italian Ministry of University and Research (MUR) under grant PRIN 2017 "EMPATHY: Empowering People in deAling with internet of THings ecosYstems" (Progetti di Rilevante Interesse Nazionale – Bando 2017, Grant 2017MX9T7H).

We thank Francesca Cerruto for supporting the research goals of this work.

References

[1] M. Antonakakis, T. April, M. Bailey, M. Bernhard, E. Bursztein, J. Cochran, Z. Durumeric, J. A. Halderman, L. Invernizzi, M. Kallitsis, D. Kumar, C. Lever, Z. Ma, J. Mason, D. Menscher, C. Seaman, N. Sullivan, K. Thomas, and Y. Zhou. Understanding the Mirai Botnet. In *Proceedings of the 26th USENIX Conference on Security Symposium*, SEC'17, page 1093–1110, USA, 2017. USENIX Association.

[2] L. Atzori, A. Iera, and G. Morabito. The internet of things: A survey. *Computer networks*, 54(15):2787–2805, 2010.

[3] B. Breve, G. Cimino, and V. Deufemia. Towards explainable security for ECA rules. In *Proceedings of the 3rd International Workshop on Empowering End-Users in Dealing with Internet of Things Ecosystems*, EMPATHY'22, 2022.

[4] B. Breve, G. Cimino, and V. Deufemia. Identifying security and privacy violation rules in trigger-action IoT platforms with NLP models. *IEEE Internet of Things Journal*, 10(6):5607–5622, 2023.

[5] B. Breve, G. Desolda, V. Deufemia, F. Greco, and M. Matera. An end-user development approach to secure smart environments. In *Proceedings of 8th International Symposium on End-User Development*, IS-EUD'21, pages 36–52. Springer, 2021.

[6] C. Cobb, M. Surbatovich, A. Kawakami, M. Sharif, L. Bauer, A. Das, and L. Jia. How risky are real users' IFTTT applets? In *Proceedings of the Sixteenth USENIX Conference on Usable Privacy and Security*, pages 505–529, 2020.

[7] F. Corno, L. De Russis, and A. Monge Roffarello. Hey-TAP: Bridging the gaps between users' needs and technology in IF-THEN rules via conversation. In *Proceedings of the International Conference on Advanced Visual Interfaces*, AVI'20, pages 1–9, 2020.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] G. Ghiani, M. Manca, F. Paternò, and C. Santoro. Personalization of context-dependent applications through trigger-action rules. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(2):1–33, 2017.

[10] T.-H. Huang, A. Azaria, O. J. Romero, and J. P. Bigham. Instructablecrowd: Creating if-then rules for smartphones via conversations with the crowd. *arXiv preprint arXiv:1909.05725*, 2019.

[11] A. Krishna, M. Le Pallec, R. Mateescu, and G. Salaün. Design and deployment of expressive and correct web of things applications. *ACM Trans. Internet Technol.*, 3(1):1–30, 2021.

[12] C. Liu, X. Chen, E. C. Shin, M. Chen, and D. Song. Latent attention for if-then program synthesis. *Advances in Neural Information Processing Systems*, 29, 2016.

[13] B. Manaris. Natural language processing: A human-computer interaction perspective. In *Advances in Computers*, volume 47, pages 1–66. Elsevier, 1998.

[14] X. Mi, F. Qian, Y. Zhang, and X. Wang. An empirical characterization of ifttt: ecosystem, usage, and performance. In *Proceedings of the 2017 Internet Measurement Conference*, pages 398–404, 2017.

[15] S. C. Mukhopadhyay and N. K. Suryadevara. *Internet of things: Challenges and opportunities*. Springer, 2014.

[16] C. Quirk, R. Mooney, and M. Galley. Language to code: Learning semantic parsers for if-this-then-that recipes. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 878–888, 2015.

[17] Q. Wang, P. Datta, W. Yang, S. Liu, A. Bates, and C. A. Gunter. Charting the attack surface of trigger-action IoT platforms. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, page 1439–1453, New York, NY, USA, 2019. Association for Computing Machinery.

[18] D. Xiao, Q. Wang, M. Cai, Z. Zhu, and W. Zhao. A3ID: an automatic and interpretable implicit interference detection method for smart home via knowledge graph. *IEEE Internet of Things Journal*, 7(3):2197–2211, 2019.

[19] Z. Yao, X. Li, J. Gao, B. Sadler, and H. Sun. Interactive semantic parsing for if-then recipes via hierarchical reinforcement learning. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019.

[20] I. N. B. Yusuf, L. Jiang, and D. Lo. Accurate generation of trigger-action programs with domain-adapted sequence-to-sequence learning. In *Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension*, ICPC '22, page 99–110, New York, NY, USA, 2022. Association for Computing Machinery.

[21] E. Zeng, S. Mare, and F. Roesner. End user security and privacy concerns with smart homes. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 65–80, Santa Clara, CA, July 2017. USENIX Association.

A Comparative Analysis of Agile Teamwork Quality Instruments in Agile Software Development: A Qualitative Approach

Ramon Santos*, Felipe Cunha*, Thiago Rique*, Mirko Perkusich*, Hyggo Almeida*
Angelo Perkusich*, Ícaro Costa†
* Intelligent Software Engineering (ISE) Group @ VIRTUS, Federal University of Campina Grande
Ícaro Costa† Fortaleza University
Emails: {ramon.santos, felipe.cunha, thiago.rique, mirko, hyggo, angelo.perkusich}@virtus.ufcg.edu.br
and psi.icarocosta@gmail.com.

***Abstract*—[Context] Multiple models (or instruments) for measuring Teamwork Quality (TWQ) for Agile Software Development (ASD) have been created. Regardless, such models have different constructs and measures, with a limited understanding of how they are related with literature factors in ASD. [Objective] Our goal is to understand how specific instruments for ASD are related, considering the relation with ASD literature factors. [Method] We analyzed three specific teamwork instruments for ASD (ASD instruments), namely ATEM, aTWQ and TWQ-BN, comparing quantitatively factors and questions to identify which ones such instruments use most and patterns among ASD literature factors. Then, we compared them qualitatively with ASD factors, given that they are specific instruments in agile context considering the solid theories that support them. [Results] The results showed that the Team Orientation and Coordination themes were identified in the first and second positions, considering the frequencies of instrument questions and literature-based Thematic Network themes (factors). Qualitative concepts can be investigated considering the ASD factors from the knowledge of the identified parts of the agile instruments. [Conclusion] There is conceptually a correlation between the identified frequencies of the ASD factors with the ASD instruments factors. We argue to add other ASD instruments to be compared to solidify the results found in this study, so we advocate further studies on this topic.**

***Index Terms*—teamwork, teamwork quality, teamwork effectiveness, Teamwork instrument, agile software development.**

I. Introduction

The success of Agile Software Development (ASD) heavily relies on the competencies, interactions, and skills of its professionals [29, 33]. As software teams are the critical source of agility in ASD [34, 11], people are a crucial resource [26, 34, 3], and the quality of team interactions can significantly impact a project's outcome. Hence, teamwork quality (TWQ) is critical for agile projects' success [20, 7, 21]. The industry is rapidly adopting ASD [31], and the need for systematic team development [25] has compelled researchers to focus on teamwork aspects increasingly.

A team can be defined as a social system of two or more people which is embedded in an organization (context), whose members perceive themselves as such and are perceived as members by others (identity), collaborating on a common task (teamwork) [1, 14, 13]. The main focus of Teamwork Quality research is on the quality of interactions within teams rather than team members' (task) activities. Starting from the widespread fundamental proposition that the success of work conducted in teams depends (beyond the quantity and correctness of the task activities) on how well team members collaborate or interact, the construct teamwork quality (TWQ) was proposed [16] as a comprehensive concept of the quality of interactions in teams. To capture the nature of team members working together, six facets of the collaborative team process integrate into the concept of TWQ: Communication, Coordination, Balance of Member Contribution, Mutual Support, Effort, and Cohesion. These facets capture both task-related and social interaction within teams. Research has shown that TWQ has a positive impact on team development [17]. Researchers argued about the importance of assessing TWQ to increase the chances of succeeding with ASD [17],[23][25].

In this context, researchers have proposed instruments for assessing teamwork quality in agile context, such as: (i) a Radar Plot [24] that considers five dimensions for assessing TWQ (Shared Leadership, Orientation, Redundancy, Learning, and Autonomy); (ii) a Structural Equation Model [20] (TWQ-SEM), based on a differentiated replication from [16], which considered that the teamwork construct is comprised of six variables: Communication, Coordination, Balance of Member Contribution, Mutual Support, Effort, and Cohesion.

All the instruments mentioned are generic (not using specific terms of ASD) and cannot represent specific situations in the agile context. Based on this finding, recently specific instruments for ASD have emerged: the aTWQ instrument [25] was developed based on the TWQ instrument [17], the ATEM instrument [32] was developed based on the Big Five theory [28], (iii) a Bayesian networks-based model (TWQ-BN) [8] was developed based on the TWQ instrument [17]. The TACT instrument [10] was developed based on the TCI instrument [2]. The STEM instrument [35] was developed considering that some specific factors in Scrum.

Silva et al. [30] performed a quantitatively comparative instruments study in ASD considering the instruments: TWQ-SEM [20] and TWQ-BN [8] instrument. However, the authors' study was conducted only from a quantitative perspective, neither investigating the instruments' questions nor providing a better understanding of how these instruments relate to each other at the question level.

Although the literature on TWQ of agile teams has evolved, there was no unified understanding of what factors influence teamwork in ASD. To better understand the factors associated with teamwork in the literature, Freire et. al. [9] developed a literature-based Thematic Network identifying the most frequent codes and themes (ASD factors) in agile teamwork literature in

10.18293/DMSVIVA2023-217

ASD. Freire et al. [9] argued that the thematic network can support their decision-making process. Practitioners can use it as a reference for understanding the factors and dimensions that comprise ASD Teamwork. With this, they can, for example, define mechanisms to monitor such dimensions and use the collected data as a reference to drive actions towards improving the team's performance.

However, for this Freire et al. [9] thematic network to have a practical use, it is necessary to identify how these codes and themes are being considered in the instruments that measure the TWQ construct in agile context. It is important to understand how these factors are associated with the factors and questions of the Agile teamwork instruments, so that they can be used in practice by teamwork instruments.

To address the research gap, we investigated current Agile teamwork Quality instruments in ASD, named from now on only "teamwork instruments", using a quantitative and qualitative approach by comparing the ASD factors and the questions for each instrument. This paper presents our findings, which represent the comparison of the current instruments in this area of research. To our knowledge, this is the first work that compares three ASD Teamwork instruments quantitatively and qualitatively at question level.

This paper is organized as follows: Section II presented the general information of the ASD Teamwork instruments compared in this work. Section III describes the employed research method. Section IV presents the results, followed by a discussion in Section V. Section VI covers the study's limitations and threats to validity. Lastly, Section VII presents our final remarks, discussing potential future work.

II. Background

In this section we presented the three ASD instruments compared in this work: ATEM, aTWQ e TWQ-BN.

ATEM - Agile teamwork effectiveness model [32]: Teamwork is crucial in software development, particularly in agile development teams which are cross-functional and where team members work intensively together to develop a cohesive software solution. Effective teamwork is not easy; prior studies indicate challenges with communication, learning, prioritization, and leadership. Nevertheless, there is much advice available for teams, from agile methods, practitioner literature, and general studies on teamwork to a growing body of empirical studies on teamwork in the specific context of ASD. The ATEM [32] model is based on evidence from focus groups, case studies, and multivocal literature and is a revision of a general Big Five [28] team effectiveness model. The ATEM [32] model is comprised of shared leadership, team mentoring, redundancy, adaptability, and peer feedback. Coordination mechanisms are needed to facilitate these components. Coordination mechanisms are shared mental models, communication and mutual trust. ATEM instrument has 31 questions.

aTWQ - Agile Team Work Quality [25]: Based on Hoegl and Gemuenden's study [17] and a systematic literature review about challenges and sucess factors for large-scale agile transformations performed by Paasivaara et al. [6]. Poth et al. [25] derived the aTWQ at initial team-level approach covering the following six factors: communication, coordination, balance of contribution, mutual support, effort, and cohesion. These six quality aspects lead to team performance [20], legitimating economically the effort for measurement and further TWQ improvement. They combined these aspects with those of TCI [2] and defined 19 related questions to come up with a holistic team evaluation questionnaire for aTWQ [25].

TWQ-BN - Teamwork Quality Bayesian networks [8] - According to the agile principles and values, as well as recent research articles, teamwork factors are critical to achieving success in agile projects. However, teamwork does not automatically arise. There are some existing instruments with the purpose of assessing the teamwork quality based on Structural Equation Modeling (i.e., empirically derived) and Radar Plot [24], but they may not be useful in a concrete situation because these techniques are not advised for prediction and diagnosis purposes. TWQ-BN instrument has 17 factors, one factor for each question.

III. Research Design

This study aims to examine, compare and synthesize the three specific instruments that measure Teamwork in ASD: ATEM [32], aTWQ [25] and TWQ-BN [8]. We used the literature-based Thematic Network codes and themes identified by Freire et. al. [9] as a basis of comparison, comparing them with three ASD Teamwork instruments factors and questions. Next, we present the study design.

A. Research questions

We aimed to perform a quantitativy and qualitative comparison between literature-based Agile Teamwork factors found by Freire et al. [9] and new Teamwork instruments factors in ASD and identify trends in this comparison by focusing on the following research questions (RQs):

- **RQ1.** How are literature-based Agile Teamwork factors (codes and themes) and ATEM, aTWQ, and TWQ-BN Agile Teamwork instruments factors and questions are quantitatively related?
- **RQ2.** How are literature-based Agile Teamwork factors (codes and themes) and ATEM, aTWQ, and TWQ-BN Agile Teamwork instruments factors and questions are qualitatively related?
- **RQ3.** How literature-based Agile Teamwork factors (codes and themes) can be investigated by researchers and practitioners with support of the instruments ATEM, aTWQ and TWQ-BN?

B. Choosing the Agile Teamwork instruments in ASD

We chose comparing the instruments ATEM [32], aTWQ [25] and TWQ-BN [8] because they are specific to the agile context. Instruments like TWQ [17] and TCI [2] are considered generic, therefore, they are outside our analysis. We did not include STEM [35] instrument in the comparison due to it being specific to Scrum nor TACT [10] because is an instrument to assess the organizational climate of agile teams, not focusing specifically in teamwork quality construct.

C. Literature-based Codes considered for the comparison with Agile Teamwork Instruments Factors and Questions

Freire et. al. [9] presented a literature-based Thematic Network identifying the following Teamwork ASD Themes and ASD codes in Table I. For example, for ASD Theme "Coordination" there are the following ASD codes: Coordination, Performance Monitoring, Task Novelty and Familiarity, and so on for the other ASD themes. In Table II are presented the factors of the ATEM, aTWQ and TWQ-BN instruments compared in this work, for each factor, there are several associated questions.

To see all Freire et.al. [9] ASD factors, questions of these instruments, and analysis, is available in the supplementary material[1].

For each Teamwork ASD code identified by Freire et al. [9], we performed a string search having the ASD code as a string word on the following ASD Instruments questions: ATEM instrument [32], aTWQ instrument [25] and TWQ-BN instrument [8]. For each Teamwork ASD code matched, we stored the question. Next, we measured the frequency of occurrence and compared the questions of these instruments aiming to give directions about how the factors (ASD code) have been used in ATEM, aTWQ and TWQ-BN instruments.

TABLE I: ASD Themes and ASD Codes in ASD identified in Freire et. al. [9] work

ASD Theme	ASD Code
Communication	Communication
Coordination	Coordination Performance Monitoring Task Novelty Familiarity
Organization Culture	Culture Structure Team Size Organization Support
Members Personality	Individual Differences Heterogeneity Personality
Management Mechanisms	Management Planning Discussion Implementation Evaluation Information Radiators Decision-Making
Team Orientation	Team Orientation Value Diversity Goals Roles Holistic Team Involvement Team Experience in the Organization Trust Motivation Norms
Expertise	Tools knowledge Collective Knowledge Adequate Skills Redudancy Team Experience with Work
Collaboration	Interdependence Collaboration
Shared Leadership	Shared Leadership Formal Leadership
Team Autonomy	Team Autonomy Task Control
Feedback	Awareness Acceptance Feedback
Team Learning	Team Learning
Communication	Communication
Cohesion	Cohesion

IV. Results

This section presents the results of this study. We compared quantitatively and qualitatively the instruments ATEM, aTWQ and TWQ-BN with the ASD codes of Freire et.al. study [9]. All the definitions of codes and themes presented in Section IV are in Freire et.al. [9] study and in the supplementary material of this work.

[1]Supplementary Material: https://figshare.com/s/13662df26088a629abf3

TABLE II: ATEM, aTWQ and TWQ-BN Instrument Factors

ATEM factor	aTWQ factor	TWQ-BN factor
TCM - Shared Mental Models	Participative safety	Teamwork
TCM - Mutual trust	Support for Innovation	Team Autonomy
TCM - Communication	Vision	Cohesion
TC - Shared leadership	Task orientation	Collaboration
TC - Peer feedback	Coordination	Self-Organizing
TC - Redundancy		Coordination
TC - Adaptability		Team Orientation
TC - Team Orientation		Communication
		Daily Meetings
		Team Distribution
		Means of Commun.
		Monitoring
		All Members Present
		Personal Attributes
		Expertise
		Shared Leadership
		Team Learning

A. Quantitative Comparison between ASD factors (Codes and Themes) and ASD Instrument Factors and Questions

The quantitative analysis was based on frequency analysis, where each word of a ASD code contained in a question of the ASD instrument was computed. In Table III, it presented the themes and codes associated with agile teamwork literature identified by Freire et.al. [9]. The ASD Theme is associated with "ASD Theme" that correspond to the general concept. In the second column, there is the column "ASD Code" that correspond to the specific ASD concept. Since all the code is associated with a theme, the notation used to ASD code that will presented in this work will be: ASD Theme - ASD code. For example, in Table III the name "Team Autonomy - Task Control" represent a ASD code where the theme is "Team Autonomy" and the code is "Task Control".

Next, we analyze the matches of the ASD codes and questions in ASD instruments shown in Table III. The notation used to instrument's question that will presented in this work will be: [Number of Question]-Model-Factor-Question. For example, the question: [17]-aTWQ-Task orientation- "Do your team colleagues provide useful ideas and practical help to enable you to do the job to the best of your abilities?" The "aTWQ" correspond to the ASD instrument; the name "Task orientation" correspond to the instrument factor and the rest correspond to the instrument question.

Note that in Table III there are codes frequencies that have more than one theme. As an example, there are the codes Personality - Individual differences (4 matches) and Personality - Trust (4 matches) identified, resulting in 8 matches in Theme "Personality" since the two referred codes belong to the Personality Team, then these frequency matches were added. We did the same process for all ASD Codes in Table III. In Table IV is presented the result of the match frequency of the previous process.

In Table III, for each instrument, we identified the following ASD code frequencies: Team Autonomy - Task Control (14 matches), Coordination - Coordination (14 matches), Shared Leadership - Shared Leadership (9 matches), Communication - Communication (6 matches), Feedback - Feedback (4 matches), Personality - Trust (4 matches), Team Orientation - Team Orientation (4 matches), Team Orientation - Goals (4 matches), Team Orientation - Planning (3 matches), Coordination - Performance Monitoring (3 matches), Team Orientation - Information Radiators (3 matches), Team Orientation - Redundancy (3

TABLE III: ASD Code Frequencies in ASD instruments

ASD Code	Instrum.	#Freq	Tot.
Team Autonomy - Task Control	ATEM	4	15
	aTWQ	7	
	TWQ-BN	4	
Coordination - Coordination	ATEM	12	14
	aTWQ	1	
	TWQ-BN	1	
Shared Leadership - Shared Leadership	ATEM	8	9
	aTWQ	0	
	TWQ-BN	1	
Communication - Communication	ATEM	3	6
	aTWQ	1	
	TWQ-BN	2	
Feedback -Feedback	ATEM	4	4
	aTWQ	0	
	TWQ-BN	0	
Personality - Trust	ATEM	3	4
	aTWQ	0	
	TWQ-BN	1	
Team Orientation - Team Orientation	ATEM	3	4
	aTWQ	0	
	TWQ-BN	1	
Team Orientation - Goals	ATEM	0	3
	aTWQ	1	
	TWQ-BN	2	
Team Orientation - Planning	ATEM	1	3
	aTWQ	2	
	TWQ-BN	0	
Coordination - Performance Monitoring	ATEM	1	3
	aTWQ	1	
	TWQ-BN	1	
Team Orientation - Information Radiators	ATEM	2	3
	aTWQ	1	
	TWQ-BN	0	
Team Orientation - Redundancy	ATEM	3	3
	aTWQ	0	
	TWQ-BN	0	
Personality - Individual differences	ATEM	2	3
	aTWQ	1	
	TWQ-BN	0	
Team Orientation - Decision-Making	ATEM	0	1
	aTWQ	0	
	TWQ-BN	1	
Expertise - Tools knowledge	ATEM	1	1
	aTWQ	0	
	TWQ-BN	0	
Expertise - Adequate Skills	ATEM	1	1
	aTWQ	0	
	TWQ-BN	0	
Expertise - Task Novelty	ATEM	0	1
	aTWQ	1	
	TWQ-BN	0	
Expertise - Structure	ATEM	0	1
	aTWQ	1	
	TWQ-BN	0	
Expertise - Roles	ATEM	1	1
	aTWQ	0	
	TWQ-BN	0	
Expertise - Motivation	ATEM	0	1
	aTWQ	1	
	TWQ-BN	0	
Collaboration - Interdependence	ATEM	0	1
	aTWQ	0	
	TWQ-BN	1	
Team Learning - Team Learning	ATEM	0	1
	aTWQ	0	
	TWQ-BN	1	

matches), Personality - Individual differences (3 matches), Team Orientation - Decision-Making (1 match), Expertise - Tools knowledge (1 match), Expertise - Adequate Skills (1 match), Expertise - Task Novelty (1 match), Expertise - Structure (1 match), Expertise - Roles (1 match), Expertise - Motivation (1 match), Collaboration - Interdependence (1 match), and Team Learning - Team Learning (1 match).

TABLE IV: Frequencies between ASD themes and Agile instrument questions

ASD Theme	Instrument	#Freq	Total
Team Orientation	ATEM	9	17
	aTWQ	4	
	TWQ-BN	4	
Coordination	ATEM	13	17
	aTWQ	2	
	TWQ-BN	2	
Team Autonomy	ATEM	4	15
	aTWQ	7	
	TWQ-BN	4	
Shared Leadership	ATEM	8	9
	aTWQ	0	
	TWQ-BN	1	
Personality	ATEM	5	7
	aTWQ	1	
	TWQ-BN	1	
Communication	ATEM	3	6
	aTWQ	1	
	TWQ-BN	2	
Expertise	ATEM	3	6
	aTWQ	3	
	TWQ-BN	0	
Feedback	ATEM	4	4
	aTWQ	0	
	TWQ-BN	0	
Collaboration	ATEM	0	1
	aTWQ	0	
	TWQ-BN	1	
Team Learning	ATEM	0	1
	aTWQ	0	
	TWQ-BN	1	
Cohesion	ATEM	0	1
	aTWQ	0	
	TWQ-BN	1	

It was identified the following ASD theme frequencies: Team Orientation (17 matches), Coordination (17 matches), Team Autonomy (14 matches), Shared Leadership (9 matches), Personality (7 matches), Communication (6 matches), Expertise (6 matches), Feedback (4 matches), Collaboration (1 match), Team Learning (1 match), and Cohesion (1 match) - as summarized in Table IV. Comparing the ASD theme frequencies in Freire et.al. [9] in Table V and the results of the Frequency Themes in the Instruments questions in Table IV, it was found that Freire's ASD Themes Team Orientation (the highest frequency with 22 matches) and Coordination (the second highest frequency with 16 matches) are the same ranking position found in this work: Team Orientation with 17 matches and Coordination with 17 matches. This result shows that the same codes identified in Freire et.al. [9] have been used in the ASD instruments (considering the number of matches).

B. Qualitative Comparison between ASD factors (Codes and Themes) and ASD Instrument Factors and Questions

For a more in-depth comparison, we compared all questions that have ASD code names in their content as showed in Table III. The purpose of this comparison is to understand how the ASD codes of Freire et.al. study [9] are addressed in the ASD instruments.

Team Autonomy - Task Control: Task Control refers to the "degree of control or authority that a team has over its internal work processes"[22]. The code Team Autonomy - Task Control has 14 matches with Instrument factors questions. The ATEM instrument has four matches in the following questions: [3]-ATEM-TCM-Shared Mental Models- "Common understanding of tasks", [10]-ATEM-TCM-Communication- "The team follows up on the progress of tasks", [25]-ATEM-TC-Redundancy-

TABLE V: ASD Theme frequencies in Freire et al. work

ASD Theme	ASD Code	#Freq	Total
Team Orientation	Orientation	7	22
	Value Diversity	1	
	Goals	2	
	Roles	2	
	Holistic Team Involvement	1	
	Experience in the Organi.	1	
	Trust	5	
	Motivation	1	
	Norms	2	
Coordination	Coordination	5	16
	Performance Monitoring	9	
	Task Novelty	1	
	Familiarity	1	
Expertise	Tools knowledge	2	15
	Collective Knowledge	4	
	Adequate Skills	1	
	Redundancy	7	
	Experience with Work	1	
Management Mechanisms	Management	4	10
	Planning	1	
	Discussion	1	
	Implementation	1	
	Evaluation	1	
	Information Radiators	1	
	Decision Making	1	
Shared Leadership	Shared Leadership	8	9
	Formal Leadership	1	
Communication	Communication	9	9
Organization Culture	Culture	4	8
	Structure	1	
	Team Size	2	
	Organization Support	1	
Collaboration	Interdependence	1	8
	Collaboration	7	
Learning	Learning	8	8
Members Personality	Individual Differences	1	5
	Heterogeneity	1	
	Personality	3	
Team Autonomy	Autonomy	4	5
	Task Control	1	
Feedback	Awareness	1	5
	Acceptance	1	
	Feedback	3	
Cohesion	Cohesion	3	3

"Completion of the whole task or parts of tasks by other team members", [29]-ATEM-TC-Team orientation- "Increased task involvement, information sharing, strategizing, and participatory goal setting". The aTWQ instrument has seven matches in the following questions: [1]-aTWQ-Participative safety- "Do we have a "we are in it together" attitude driven by the ability and willingness to help and support each other in carrying out their tasks?", [12]-aTWQ-Support for Innovation- "Do team members provide practical support for new ideas and their application by prioritizing the teams' task over other obligations?", [17]-aTWQ-Task orientation- "Do your team colleagues provide useful ideas and practical help to enable you to do the job to the best of your abilities?", [18]-aTWQ-Task orientation- "Are team members prepared to question the basis of what the team is doing?", [19]-aTWQ-Task orientation- "Does the team critically appraise potential weaknesses in what it is doing in order to achieve the best possible outcome?", [20]-aTWQ-Task orientation- "Do members of the team build on one another's ideas in order to achieve the highest possible standards of performance?", [21]-aTWQ-Coordination- "Is there a common understanding when working on parallel subtasks, and agreement on common work breakdown structures, schedules, budgets, and deliverables?". In the TWQ-BN instrument, has four matches: [2]-TWQ-BN-Team Autonomy- "There is no external agent interfering on how the team executes its tasks. The external agent collaborates with them to define what will be", [6]-TWQ-BN-Coordination- "The team executes its tasks in a synchronous and integrated manner", [12]-TWQ-BN-Monitoring- "The team members expose their obstacles and progress regarding their tasks in a clear and objective way" and [15]-TWQ-BN-Expertise- "The team members have the necessary knowledge for developing the tasks with redundancy." Analyzing these questions matches, it is possible to say that all, in general, are associated with the team autonomy, therefore, one can envision a conceptualization joining the Task Control factors with those identified in the questions of the instruments: ATEM-TCM-Shared Mental Models, ATEM-TCM-Communication, ATEM-TC-Team orientation, aTWQ-Participative safety, aTWQ-Support for Innovation, aTWQ-Task orientation, aTWQ-Coordination, TWQ-BN-Team Autonomy, TWQ-BN-Coordination, TWQ-BN-Monitoring and TWQ-BN-Expertise. Analyzing the matches, it is suggested that a team that has autonomy in its tasks and work processes, probably, has a shared mental models, a team orientation, a participative safety, a support for innovation, a task orientation, a coordination, a team autonomy, a monitoring and a expertise.

Coordination - Coordination: Coordination refers to team members executing their activities in a timely and integrated manner. It implies that the performance of some team members influences the performance of others. This may involve an exchange of information that subsequently influences another member's performance [23]. The degree of common understanding regarding the interrelatedness and status of individual contributions [16]; It refers to team members executing their activities in a timely and integrated manner and it is linked to the performance of teams [12]. The code Coordination has 14 matches with Instrument Factors questions. In ATEM instrument, there are Team Coordination Mechanisms (TCM) and Teamwork Components (TC). The TCM are composed by: Shared Mental Models, Mutual Trust, and Communication, showing how important the Coordination Mechanisms are to Agile Teamwork Effectiveness. The TC are composed by: Shared leadership, Redundancy, Peer feedback, Adaptability, and Team Orientation. The questions in ATEM-TCM are general, such as: [2]-ATEM-TCM-Shared Mental Models- "Common understanding of goals", [3]-ATEM-TCM-Shared Mental Models- "Common understanding of tasks",[7]-ATEM-TCM-Mutual trust- "Information sharing". In the aTWQ instrument, there is only one question referred to "Coordination": [21]-aTWQ-Coordination-"Is there a common understanding when working on parallel subtasks, and agreement on common work breakdown structures, schedules, budgets and deliverables?". Note that in this question, the aTWQ instrument aggregates several characteristics of coordination factors into a single question. In the TWQ-BN instrument, there is only one question too: [6]-TWQ-BN-Coordination- "The team executes its tasks in a synchronous and integrated manner." Note that in this question, the TWQ-BN instrument brings a more generic concept of coordination. Analyzing these matches, it is possible to say that all, in general, are associated with Coordination that following mechanisms that TCM and TC in ATEM instrument. These factors represent a large number of factors, including: Shared Mental Models, Mutual Trust, and Communication, showing how important the Coordination Mechanisms are to Agile Teamwork Effectiveness. It is suggested by the high frequency that

Coordination is one of the most important factors in Teamwork quality. The ATEM instrument defines the Coordination factor with a greater completeness while the aTWQ and TWQ-BN instruments are more generic. As recommendation, for more in-depth analysis, we recommend using the ATEM instrument when analyzing the Team Coordination.

Shared Leadership - Shared Leadership: Leadership is rotated to the person with key knowledge, there is jointly shared decision authority [27]. The code Shared Leadership has nine matches with Instrument Factors questions: [13]-ATEM-TC-Shared leadership- "The agile team facilitates team problem-solving", '[14]-ATEM-TC-Shared leadership- "The agile team determines performance expectations and acceptable interaction patterns", [15]-ATEM-TC-Shared leadership- "The agile team synchronizes and combines individual team member contributions using agile practices combined with automated tools", [16]-ATEM-TC-Shared leadership- "The agile team seeks and evaluates information that affects team functioning", [17]-ATEM-TC-Shared leadership- "Agile values and methodologies determine team member roles", [18]-ATEM-TC-Shared leadership- "Agile values and methodologies determine the frequency and type of preparatory meetings and feedback sessions", [19]-ATEM-TC-Shared leadership- "A servant leader facilitates a boundary-spanning function", [20]-ATEM-TC-Shared leadership-"Agile team practices provide a planning function". In aTWQ there is no specific question related to "Shared Leadership", in TWQ-BN instrument there is one question:"[16]-TWQ-BN-Shared Leadership- "The decision authority and leadership is shared." Shared Leadership code has a great importance in ATEM instrument with eight matches and is one of the most important factors for Teamwork Quality in Agile Context. In TWQ-BN instrument there is one generic question. The ATEM instrument has a greater completeness in Shared Leadership code.

Communication - Communication: Communication provides a means for the exchange of information among team members [16]. The fundamental component of teamwork is communication. It provides a means to exchange information, share ideas among team members, coordinate efforts and provide feedback [36]. The code Communication has six matches with Instrument Factors questions. In the ATEM instrument, we found more generic questions, for example: [10]-ATEM-TCM-Communication- "The team follows up on the progress of tasks", [11]-ATEM-TCM-Communication- "Visualize project information" and [12]-ATEM-TCM-Communication- "Facilitate informal communication". In aTWQ instrument, there is not explicitly a communication factor, but in question [4]-aTWQ-Participative safety- "Do people keep each other informed about work-related issues in the team supported by a frequent communication?". In TWQ-BN instrument, two questions are related to "Communication": [8]-TWQ-BN-Communication- "The communication is effective" and [11]-TWQ-BN-Means of Communication- "The team members communicate face-to-face whenever possible". Communication factor have a great importance in the three instruments: in ATEM, the questions are associated with informal communication and visualization of project information. In aTWQ instrument, it is considered that for a team to stay informed about work matters, good communication is necessary. In TWQ-BN instrument, highlights the importance of effective, face-to-face communication whenever possible. For a team to coordinate the tasks, it must communicate. Since agile is based on tacit knowledge sharing [4], Communication is a must factor to assess agile teams [8] [18], Furthermore, in agile context, daily meetings play an important role on synchronizing the team members' tasks, as well as removing impediments and mitigating risks.

Feedback - Feedback: Feedback involves the giving, seeking, and receiving of information among team members. Giving feedback refers to providing information regarding other members' performance. Seeking feedback refers to requesting input or guidance regarding performance and to accepting positive and negative information regarding performance, e.g. * responding to other members' requests for information about their performance * accepting time-saving suggestions offered by other team members" [23]. "It involves providing information regarding other members' performance, requesting input or guidance regarding performance of self and to accept positive and negative information regarding performance" [12]."High performance teams also get constant feedback on their productivity and effectiveness both internally and from external resources, and use this feedback to make improvements to the group work" [11]. The code Feedback has four matches with Instrument Factors questions. [8]-ATEM-TCM-Mutual trust- "Willingness to admit mistakes and accept feedback"; [18]-ATEM-TC-Shared leadership- "Agile values and methodologies determine the frequency and type of preparatory meetings and feedback sessions"; [21]-ATEM-TC-Peer feedback- "Identifying mistakes and lapses in other team members' actions"; [22]-ATEM-TC-Peer feedback- "Regular feedback regarding team member actions to facilitate self-correction". Analyzing the frequencies, the ATEM instrument is the only instrument that talks about the importance of feedback associating this code with the factors: ATEM-TCM-Mutual trust, ATEM-TC-Shared leadership and ATEM-TC-Peer feedback.

Personality - Trust: Without sufficient trust, team members will extend time and energy protecting, checking and inspecting each other as opposed to collaborating to provide value-added ideas [19]. They understand that agility depends on trusting individuals to apply their competency in effective ways [5]. The code Personality - Trust has four matches with Instrument Factors questions: [7]-ATEM-TCM-Mutual trust- "Information sharing"; [8]-ATEM-TCM-Mutual trust- "Willingness to admit mistakes and accept feedback"; [9]-ATEM-TCM-Mutual trust- "Supportive team social climate"; [7]-TWQ-BN-Team Orientation- "The team members trust each other and feel motivated to work together for achieving the team goals." Analyzing the frequencies, the ATEM instrument has a specific factor related to the Trust code called "ATEM-TCM-Mutual trust" and the TWQ-BN instrument has a factor called "TWQ-BN-Team Orientation". The ATEM instrument associated trust code with "Information sharing", "Willingness to admit mistakes and accept feedback" and the TWQ-BN instrument has a more general question. So, if the company needs more detailed information for the purposes of diagnosing the situation of the teams about the trust of team members, it is more recommended to apply the ATEM instrument.

Team Orientation - Team Orientation: Team orientation refers to the team tasks and the attitudes that team members have towards one another. It reflects an acceptance of team norms, the level of group cohesiveness, and the importance of team membership, e.g.-assigning high priority to team goals and participating willingly in all relevant aspects of the team [23].

Refers to belief of team members in the importance of team goals over individual member goals, propensity to take other's behavior into account during group interaction. It reflects an acceptance of team norms, the level of group cohesiveness and the importance of team membership" [12]. The ability to take other team member's behavior into account and set team goals over individual goals [15]. The code Team Orientation has three matches with ATEM Instrument Factors questions and one match with TWQ-BN instrument factor: [28]-ATEM-TC-Team orientation- "Increased task involvement, information sharing, strategising, and participatory goal setting"; [30]-ATEM-TC-Team orientation - "Increased task involvement, information sharing, strategising, and participatory goal setting"; [31]-ATEM-TC-Team orientation- "The team sticks together and remains united"; [7]-TWQ-BN-Team Orientation- "The team members trust each other and feel motivated to work together for achieving the team goals." Analyzing the frequencies, the code Team Orientation, this is a factor in ATEM with 3 questions and a factor in TWQ-BN with one question. The Team Orientation is related with control mechanisms in ATEM what is associated with achieve the goals.

Team Orientation - Goals: That members are clear about team goals is the single most important part of high performing teams [11]. The code Team Orientation-Goals has one match in [6]-aTWQ-Participative safety- "Do we keep in touch with one another as a team by accepting that team goals are more important than individual goals? It is suggestive a relationship between Team Orientation - Goals code and Participative safety, considering that the "team goals are more important than individual goals".

Team Orientation - Planning: "The best teams spend time planning how they will solve problems and make decisions [11]". The code Team Orientation-Planning has one match in ATEM instrument: [20]-ATEM-TC-Shared leadership- "Agile team practices provide a planning function" and two matches in aTWQ instrument: [5]-aTWQ-Participative safety-"- "Is there a lot of give and take by the team members' motivation to maintain the team? - Innovation and Planning Iteration-IV"; [10]-aTWQ- "Support for Innovation - Do people in this team always search for fresh, new ways of looking at problems-Innovation and Planning Iteration-IV?" It is suggestive that Team Orientation - Planning code is associated with ATEM-TC-Shared leadership when say "Agile team practices provide a planning function".

Team Orientation - Performance Monitoring: "It is the ability to develop a common understanding of the team environment through observing the activities of other team members and apply appropriate task strategies to accurately monitor teammate performance to recognize when a team member performs correctly [12]. The code Team Orientation - Performance Monitoring has three matches with Instrument Factors questions: [14]-ATEM-TC-Shared leadership- "The agile team determines performance expectations and acceptable interaction patterns"; [20]-aTWQ-Task orientation- "Do members of the team build on one another's ideas in order to achieve the highest possible standards of performance?"; [12]- TWQ-BN- Monitoring- "The team members expose their obstacles and progress regarding their tasks in a clear and objective way." Analyzing the frequencies, the code Team Orientation - Performance Monitoring are found in three instruments evidencing the importance for teamwork quality.

Team Orientation - Information Radiators: "Information radiators, such as burn charts, allow teams to clearly visualize current project status and what is required to complete goals. Such information radiators were discussed as being invaluable sources of motivation, excitement, and team cohesion" [37]. The code Team Orientation - Information Radiators has three matches: two matches with ATEM Instrument and one match with aTWQ instrument: [16]-ATEM-TC-Shared leadership- "The agile team seeks and evaluates information that affects team functioning"; [28]-ATEM-TC-Team orientation- "Increased task involvement, information sharing, strategising, and participatory goal setting"; [4]-aTWQ-"Participative safety - Are there real attempts to share information throughout the team driven by openness of the information exchange?" Analyzing the frequencies, the code Team Orientation - Information Radiators, it's suggestive the information is important for team functioning.

Team Orientation - Redundancy: Redundancy is associated with members that have multiple skills so that they can perform (parts of) each others' tasks [27]. It was found three matches: [23]-ATEM-TC-Redundancy- "Recognition by potential backup providers that there is a workload distribution problem in their team". [24]-ATEM-TC-Redundancy- Shifting of work responsibilities to underutilized team members", [25]-ATEM-TC-Redundancy- "Completion of the whole task or parts of tasks by other team members". Analyzing the frequencies, the only instrument that matches with Team Orientation - Redundancy was ATEM that has one specific factor for this. So, the Redundancy is associated with "Recognition by potential backup providers that there is a workload distribution problem in their team" and "Shifting of work responsibilities to underutilized team members" and "Completion of the whole task or parts of tasks by other team members" [32].

Personality - Individual Differences: Successful teams also accept differences in people as long as their behavior helps task accomplishment [11]. The code Personality - Individual Differences has three matches with Instrument Factors questions: [6]-ATEM-TCM-Shared Mental Models- "Common understanding of individual skills and expertise"; [15]-ATEM-TC-Shared leadership- "The agile team synchronizes and combines individual team member contributions using agile practices combined with automated tools"; [6]-aTWQ-Participative-safety- "Do we keep in touch with one another as a team by accepting that team goals are more important than individual goals?"; Analyzing the frequencies, the Individual Differences are more associated with Participative safety considering that team accepting that team goals are more important than individual goals.

Team Orientation - Decision-Making: It does not matter if the decision-making strategy is consensus or majority, etc., it is only important that the rules of engagement are defined beforehand [11]. Analyzing the frequencies, the only instrument that matched was [16]-TWQ-BN- Shared Leadership- "The decision authority and leadership is shared.". It was suggestive a relationship between Decision-Making and Shared leadership. The literature consider that the agile team needs a Shared Leadership [32] [35].

Expertise - Tools knowledge: There was one match in Expertise - Tools knowledge: [15]-ATEM-TC-Shared leadership- "The agile team synchronizes and combines individual team member contributions using agile practices combined with automated tools". It is suggestive and proven in the literature that agile practices with automated tools is a important factor for

teamwork quality in agile context [32], [35].

Expertise - Adequate Skills: "Refers to the required skills that the software development team must possess to execute their tasks. It is measured through the team member's perspective about the adequacy of the team competencies" [22]. There is one match in ATEM instrument: [6]-ATEM-TCM-Shared Mental Models- "Common understanding of individual skills and expertise". It is suggested that a common understanding of individual skills and expertise influenced positively the teamwork quality [32].

Expertise - Task Novelty: If the task novelty is low, it is likely that the teams have developed sufficient meta knowledge to adequately assign tasks to team members [22]. It has one match in the question: [8]-aTWQ-Support for Innovation- "Is this team always moving towards the development of new answers?". We consider a team that is looking for new answers as well as one that has knowledge of new tasks. So, it suggestive that task novelty is associated with teamwork quality.

Expertise - Structure: "The structure of the team is important. The team members must all contribute, and therefore, a successful team only consists of the smallest number of members necessary to reach the group goal. The group must also allow subgroups to form to work on smaller chores. These subgroups are not seen as a threat to the group, but as necessary and valued for their contribution to the team [11]". There is one match in aTWQ instrument: [21]-aTWQ-Coordination- "Is there a common understanding when working on parallel subtasks, and agreement on common work breakdown structures, schedules, budgets and deliverables?". This match did not have a semantic correspondence.

Expertise - Roles: "After the goal is defined the group can get organized and decide what needs to be done and who does what. The most important thing is that each member really knows what their role is, independently of if they volunteered for the role or not, i.e., both the expectations and the process need to be clear [11]." There is one match with "[17]-ATEM-TC-Shared leadership- "Agile values and methodologies determine team member roles". It is possible to say that roles in agile teams are determined by agile values and methodologies [32].

Expertise - Motivation: "There seems to be value, therefore, in frequent signs of progress towards collective goals. Such indicators were seen to strongly support individual motivation to contribute to team efforts [37]. There is one match in aTWQ instrument: [5]-aTWQ-Participative safety- "Is there a lot of give and take by the team members' motivation to maintain the team?". Motivation was associated with the quality of a team's work [17].

Collaboration - Interdependence: "In high performance teams, the tasks demand that members work together as a unit or in subgroups to reach the goal [11]." There is one match in TWQ-instrument: [4]-TWQ-BN-Collaboration- "There is a high degree of collaboration in the team for achieving success on the project development." Is suggestive that the existence of high degree of collaboration is associated with success on the project development [8].

Team Learning - Team Learning: "It involves the ability to identify the changes in the team environment and adjust the strategies as needed" [12]. The instrument TWQ-BN has one match: [17]-TWQ-BN-Team Learning- "The team adapts itself to changes in the team environment and adjust the strategies as needed.". The instruments aTWQ and STEM don't explicitly talk about team learning, but has the "Team Learning" in other factors.

Cohesion - Cohesion: Team cohesion refers to the degree to which team members desire to remain on the team [16]. The TWQ-BN instrument has one match: [3]-TWQ-BN-Cohesion- "The team works cohesively and synchronously, prioritizing the team goals, and self-organize efficiently." Cohesion is one important factor in [17].

C. Responses to research questions

This section introduces a discussion on research questions, trends observed, attributes, and data collection mechanisms.

RQ1. How are literature-based Agile Teamwork factors (codes and themes) and ATEM, aTWQ, and TWQ-BN Agile Teamwork instruments factors and questions are quantitatively related?

We mapped the factors of the three instruments (ATEM, aTWQ, and TWQ-BN), then we compared them with ASD codes and themes found by Freire et al. [9]. The objective is to understand how the ASD factors (codes and themes) and instrument factors and questions are related. Then, we intended to identify trends in these factors. We noted that the codes with more Teamwork instrument questions are Team Autonomy - Task Control (14 questions matched), Coordination - Coordination (14 questions matched), and Shared Leadership - Shared Leadership (9 matches). Considering the Themes analysis in Section IV-A, the result of this work confirmed Freire et.al. [9] results in which the first two frequencies of these studies are in the same order: Team Orientation and Coordination.

RQ2. How are literature-based Agile Teamwork factors (codes and themes) and ATEM, aTWQ, and TWQ-BN Agile Teamwork instruments factors and questions are qualitatively related? From the Section IV-B it's possible to say that the instruments ATEM, aTWQ and TWQ-BN brought in these instruments questions new concepts directly associated with the agile context, among them: agile practices, daily sprints, retrospective meetings, etc. Thus, the present work demonstrated this conceptual evolution of the ASD terms in Freire et.al. [9] work. As examples: [15]-ATEM-TC-Shared leadership- "The agile team synchronizes and combines individual team member contributions using agile practices combined with automated tools"; [18]-ATEM-TC-Shared leadership- "Agile values and methodologies determine the frequency and type of preparatory meetings and feedback sessions"; It can be understood by the behaviours markers that consider agile practices[32].

RQ3. How literature-based Agile Teamwork factors (codes and themes) can be investigated by researchers and practitioners with support of the instruments ATEM, aTWQ and TWQ-BN?

The researchers can investigate whether high or lower frequencies are in fact more or less important for the teamwork quality. In this way, researchers will already have prior knowledge of which parts of the instruments to use. From the results from RQ2, it was found the frequency of appearance of each factor related to the teamwork quality and the number of corresponding questions for each instrument. With this knowledge, this work can support other works that need to use a ASD teamwork instrument for a specific purpose. As example, if a researcher needs to investigate the relationship between Feedback and Team Autonomy in a company, he can choose specific parts of ASD instruments identified in this work.

Qualitative concepts can be investigated in future works that aim to investigate the ASD factors from the knowledge of the identified parts of the agile instruments.

V. Discussion

In Section IV, we compared ASD codes found in Freire et.al. [9] with all the questions of the ASD instruments, addressing the RQs in section III-A. In Section V-B, we discuss implications for research and practice.

A. Comparison of Literature-based Teamwork Factors (codes and themes) and Teamwork Instrument Factors and Questions

We used the literature-based Thematic Network codes identified by Freire et al. [9] as a comparison base because it is the most current work that analyzed the most recurrent factors in agile teamwork works in the literature, providing evidence that these factors are important for the teamwork quality. We observed that, considering the four themes with more matches as showed in Section IV: Task Orientation (17 matches), Coordination (17 matches), Team Autonomy (14 matches), and Shared Leadership (9 matches). From the result of the analysis of the frequencies of the instruments, it is suggestive to say that: "Agile times that have task orientation, coordination, time autonomy and shared leadership are more likely to have a high teamwork quality."

B. Summary of findings

This work can support other works that need to use a ASD Teamwork Instrument for a specific purpose. As example, if a researcher needs to investigate the relationship between Feedback and Team Autonomy, he can choose what parts of instruments use. For Feedback code, there are four questions in ATEM instrument. For Team Autonomy, there are four questions in ATEM instrument, six questions in aTWQ instrument, and four questions in TWQ-BN instrument. This work highlights that the ASD literature themes: Team Orientation (14 matches), Coordination (17 matches), Team Autonomy (14 matches), and Shared Leadership (9 matches) are the most used in ASD Teamwork Instruments. We observed that, considering the four themes with more matches as showed in Section IV, we have a pattern considered that in Freire et.al. [9], the two ASD themes with more frequency were Team Orientation, and Coordination. This is an important result, as it confirms that the factors identified by Freire et. al. [9] are, in fact, those that are being used more frequently in specific ASD teamwork instruments, which were developed based on strong literature theories and empirical studies. Additionally, we compared the referred questions in the ATEM, aTWQ and TWQ-BN instruments. We noted that finding a standard terminology for ASD Teamwork factors remains challenging, and there is a need for further investigation into this area. Finally, practitioners can benefit from the study's findings by better understanding the recent Agile Teamwork instruments in ASD.

VI. Limitations and threats to validity

The results of this study may have been impacted by the frequency analisys methods based only on syntatic aspects and incompleteness of the ASD Teamwork instrument.

Quantitative analysis based only on syntactic aspects. The quantitative analysis was based on frequency analysis, where each word of a ASD code contained in a question of the ASD instrument will be computed without considering semantic aspects. This can lead to some problems, such as incorrect semantic counts, but in order for the study to have a more reproducible method, we preferred to adopt this choice. As future work, we intend to consider semantic aspects in quantitative analysis.

Incompleteness of the ASD Teamwork instruments. Three instruments were chosen for the agile context, possibly the results could be different if more instruments were added. For reasons of time and complexity of the work, at the moment, only three instruments were considered. As future works, we intend to compare more ASD instruments.

VII. Conclusions

Our study makes several contributions to the teamwork quality literature aiming to give directions for an understanding of how ASD literature-based codes and themes identified by Freire et al. [9] and Agile Instruments factors and questions in ASD are related. We identified and compared three instruments specific for ASD, showing the frequency of the matches. Further, we identified ASD Instruments questions related to ASD literature-based codes that can support other works that investigate the relationship between ASD factors by providing knowledge of specific parts of ASD instruments. In this way, researchers will be able to have greater coverage in their investigations. This study can support other studies that can increase the body of knowledge by allowing an update of the literature-based Thematic Network developed by Freire et. al. [9].

Our findings show that many factors have been used by researchers to measure teamwork quality in ASD. Also, the analyzed instruments have similar questions with different names, pointing to the need for terminology standardization. Our results can support a unified Teamwork instrument in ASD, considering the most frequent questions of each instrument.

This paper presents a comprehensive view of comparing teamwork instruments qualitatively in ASD. This study has identified new trends that should be taken into account for further research in the field. Furthermore, more investigation is still needed into comparing teamwork instruments qualitatively.

References

[1] Clayton P Alderfer. *An intergroup perspective on group dynamics.* Tech. rep. Yale Univ New Haven CT School of Organization and Management, 1983.

[2] Neil Anderson and Michael A West. "The Team Climate Inventory: Development of the TCI and its applications in teambuilding for innovativeness". In: *European Journal of work and organizational psychology* 5.1 (1996), pp. 53–66.

[3] Phillip G Armour. "The spiritual life of projects". In: *Communications of the ACM* 45.1 (2002), pp. 11–14.

[4] Barry W Boehm and Richard Turner. *Balancing agility and discipline: A guide for the perplexed*. Addison-Wesley Professional, 2004.

[5] Alistair Cockburn and Jim Highsmith. "Agile software development, the people factor". In: *Computer* 34.11 (2001), pp. 131–133.

[6] Kim Dikert, Maria Paasivaara, and Casper Lassenius. "Challenges and success factors for large-scale agile transformations: A systematic literature review". In: *Journal of Systems and Software* 119 (2016), pp. 87–108.

[7] Torgeir Dingsøyr et al. “Team performance in software development: research results versus agile principles”. In: *IEEE software* 33.4 (2016), pp. 106–110.

[8] Arthur Freire et al. “A Bayesian networks-based approach to assess and improve the teamwork quality of agile teams”. In: *Information and Software Technology* 100 (2018), pp. 119–132.

[9] Arthur Freire et al. “Towards a comprehensive understanding of agile teamwork: A literature-based thematic network”. In: SEKE. 2021.

[10] Eliezer Goncalves et al. “TACT: An insTrument to Assess the organizational ClimaTe of agile teams-A Preliminary Study”. In: *Journal of Software Engineering Research and Development* 9 (2021), pp. 18–1.

[11] Lucas Gren, Richard Torkar, and Robert Feldt. “Group Maturity and Agility, Are They Connected?–A Survey Study”. In: *2015 41st Euromicro Conference on Software Engineering and Advanced Applications*. IEEE. 2015, pp. 1–8.

[12] Chaitanya Gurram and Srinivas Goud Bandi. *Teamwork in distributed agile software development*. 2013.

[13] Richard A Guzzo and Gregory P Shea. “Group performance and intergroup relations in organizations.” In: (1992).

[14] JR Hackman. “The design of work teams. In JW Lorsch (Ed.), Handbook of Organizational Behavior, Englewood Cliffs, NJ: Prentice-Hall”. In: (1987).

[15] Lars Martin Riiser Haraldsen. “An investigation of team effectiveness in agile software development”. MA thesis. Institutt for datateknikk og informasjonsvitenskap, 2012.

[16] Martin Hoegl and Hans Georg Gemuenden. “Teamwork quality and the success of innovative projects: A theoretical concept and empirical evidence”. In: *Organization science* 12.4 (2001), pp. 435–449.

[17] Martin Hoegl and Hans Georg Gemuenden. “Teamwork quality and the success of innovative projects: A theoretical concept and empirical evidence”. In: *Organization science* 12.4 (2001), pp. 435–449.

[18] Andreas Johansson. “Toward improvements of teamwork in globally distributed agile teams”. In: (2014).

[19] YAVUZ Kozak. “Barriers against better team performance in agile software projects”. In: *Chalmers University of Technology, Sweden* (2013).

[20] Yngve Lindsjørn et al. “Teamwork quality and project success in software development: A survey of agile development teams”. In: *Journal of Systems and Software* 122 (2016), pp. 274–286.

[21] Lina Lukusa et al. “Teamwork and project success in agile software development methods: A case study in higher education”. In: *Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality*. 2020, pp. 885–891.

[22] George Marsicano et al. “The Teamwork Process Antecedents (TPA) questionnaire: developing and validating a comprehensive measure for assessing antecedents of teamwork process quality”. In: *Empirical Software Engineering* 25 (2020), pp. 3928–3976.

[23] Nils Brede Moe, Torgeir Dingsøyr, and Tore Dybå. “A teamwork model for understanding an agile team: A case study of a Scrum project”. In: *Information and software technology* 52.5 (2010), pp. 480–491.

[24] Nils Brede Moe, Torgeir Dingsøyr, and Emil A Røyrvik. “Putting agile teamwork to the test–an preliminary instrument for empirically assessing and improving agile software development”. In: *Agile Processes in Software Engineering and Extreme Programming: 10th International Conference, XP 2009, Pula, Sardinia, Italy, May 25-29, 2009. Proceedings 10*. Springer. 2009, pp. 114–123.

[25] Alexander Poth, Mario Kottke, and Andreas Riel. “Evaluation of agile team work quality”. In: *Agile Processes in Software Engineering and Extreme Programming–Workshops: XP 2020 Workshops, Copenhagen, Denmark, June 8–12, 2020, Revised Selected Papers 21*. Springer. 2020, pp. 101–110.

[26] Abirami Radhakrishnan et al. “The impact of project team characteristics and client collaboration on project agility and project success: An empirical study”. In: *European Management Journal* 40.5 (2022), pp. 758–777.

[27] Mats Angermo Ringstad, Torgeir Dingsøyr, and Nils Brede Moe. “Agile process improvement: diagnosis and planning to improve teamwork”. In: *Systems, Software and Service Process Improvement: 18th European Conference, EuroSPI 2011, Roskilde, Denmark, June 27-29, 2011. Proceedings 18*. Springer. 2011, pp. 167–178.

[28] Eduardo Salas, Dana E Sims, and C Shawn Burke. “Is there a “big five” in teamwork?” In: *Small group research* 36.5 (2005), pp. 555–599.

[29] Pedro Serrador and Jeffrey K Pinto. “Does Agile work?—A quantitative analysis of agile project success”. In: *International journal of project management* 33.5 (2015), pp. 1040–1051.

[30] Manuel Silva et al. “A Comparative Analysis of Agile Teamwork Quality Measurement Models”. In: *Journal of Communications Software and Systems* 18.2 (2022), pp. 153–164.

[31] Stavros Stavru. “A critical examination of recent industrial surveys on agile method usage”. In: *Journal of Systems and Software* 94 (2014), pp. 87–97.

[32] Diane Strode, Torgeir Dingsøyr, and Yngve Lindsjorn. “A teamwork effectiveness model for agile software development”. In: *Empirical Software Engineering* 27.2 (2022), p. 56.

[33] Dothang Truong and Thawatchai Jitbaipoon. “How can agile methodologies be used to enhance the success of information technology projects?” In: *International Journal of Information Technology Project Management (IJITPM)* 7.2 (2016), pp. 1–16.

[34] Marcel F Van Assen. “Agile-based competence management: the relation between agile manufacturing and time-based competence management”. In: *International Journal of Agile Management Systems* 2.2 (2000), pp. 142–155.

[35] Christiaan Verwijs and Daniel Russo. “A theory of scrum team effectiveness”. In: *ACM Transactions on Software Engineering and Methodology* (2023).

[36] Emily Weimar et al. “Towards high performance software teamwork”. In: *Proceedings of the 17th International Conference on Evaluation and Assessment in Software Engineering*. 2013, pp. 212–215.

[37] Elizabeth Whitworth and Robert Biddle. “The social nature of agile teams”. In: *Agile 2007 (AGILE 2007)*. IEEE. 2007, pp. 26–36.

The Influence Of Technological Factors On Dark Web Marketplace Closure

Michael Kyobe and Hishaam Damon

Abstract—**The Dark Web serves as a platform to enable a host of illegal cyber activities. One such activity is Dark Web Marketplaces that operate as e-commerce websites but facilitate the sale of illicit goods and services. Various government and law enforcement agencies have surged many resources in trying to reduce dark web marketplace-related cybercrime. Still, dark web users can set up new marketplaces that become even more demanding to infiltrate. This study aimed to understand the influence of technological factors on dark market closures and how this could aid government and law enforcement in responding to dark web marketplace challenges quicker. Literature was synthesized to identify key technological factors that influence marketplace operations. These were: Anonymization, cryptocurrencies, decentralization, and codebase. A conceptual model was then developed and analyzed using quantitative data compiled from 87 dark web marketplaces. The findings suggest each of the technological factors identified has a low likelihood of influencing marketplace closures.**

Keywords; Dark Web Marketplaces, Cybercrime, Online Anonymity, Law Enforcement, Technological Factors

I. INTRODUCTION

There are three layers to the internet. The first layer, the Surface Web, has been extensively crawled and indexed and is accessed through common browsers such as Google, Firefox, and Microsoft Edge [1]. Despite its assumed size, the surface web only accounts for approximately five percent of all the information accessible on the World Wide Web. The other ninety-five percent of information is situated on the second layer, commonly referred to as the Deep Web [2]. The Deep Web has not been extensively crawled or indexed by search engines, such as Google, meaning all the information on it is inaccessible to the public and can only be accessed by navigating to a specific internet address [3]. The Deep Web contains primary harmless and protected data such as a university intranet system or information from password-protected websites such as banking details [4].

Growing expeditiously within the Deep Web is the third and final layer to the internet known as the Dark Web or, as some refer to it, the Darknet (1). A unique web browser is required for users to access the Dark Web. The most popular of these web browsers is The Onion Router (TOR) which provides anonymity to its users through redirecting internet traffic [5]. Various illegal activities take place on the Dark Web. The present study focuses on Dark Web Marketplaces, also referred to as Dark Marketplaces and the technological factors that enable them.

A significant concern regarding Dark Web Marketplaces is surrounding regulation. As the Dark Web is part of the world wide web, which is information sharing across multiple borders, it is difficult for specific governments to regulate international activity. In addition, because of the wide range of use cases that the deep web facilitates, simply restricting access is unfeasible (6). Much to the dismay of law enforcement, the closure of Silk Road, the largest Dark Web marketplace to date, had little to no effect on curbing cybercrime, and the economy of the Dark Web (7) Dark Web marketplace revenue was estimated to have increased by two hundred million USD since 2019, going from 1.3 billion USD to 1.5 billion USD in 2021. Dark Web marketplaces are anticipated to become more user-friendly and inventive, and the marketplace aspect is expected to increase as customer demand increases (8). Darknet users continue to establish new marketplaces that become more challenging to penetrate (9). (8) stated that the Darknet would become even more problematic to infiltrate as technology advances.

Thus, the purpose of this research is to determine the influence of technological factors on dark market operations, describe better how dark markets operate and allow for formulation of appropriate regulations and assist in the broader strategy of trying to detect, intercept and respond to illegal dark market activity (1).

II. LITERATURE REVIEW

Dark Web Marketplaces are built off the idea of eCommerce and function like the Alibaba Group or eBay but differ in the strong anonymity they offer their users. This vital anonymity aspect can be attributed to the web browsers needed to access these marketplaces and the cryptocurrencies that finance transactions [10]. Dark Web marketplaces offer an extensive range of products, the most frequent being illegal goods such as drugs, malware, and weapons [11]. The anonymity component provided with Dark Web marketplaces is used to elude law enforcement [12]. There are currently forty-four (44) Dark Web marketplaces active as of the beginning of 2021. Some of these marketplaces include the third installment of the original Silk Road, called Silk Road 3.1, and the DarkFox Market, which is currently one of the largest marketplaces in 2021 [14]. In addition, since the covid-19 pandemic, Dark Web marketplaces have witnessed an increase in bulk buying from users and the sale of personal protective equipment (PPE) and other medical goods [15]. them. However, in the context of this paper, online anonymity will extend to the technological aspect in which online activity cannot be linked to an Internet Protocol (IP) address [17]. Simply hiding your identity over the internet does not ensure anonymity from Internet Service Providers (ISP's) [18].

In terms of Dark Web and Dark Web Marketplaces, The Onion Router (TOR) is a popular tool that enables anonymity from ISP's while browsing on the Dark Web [17]. Cryptocurrencies are also a critical technology that allows anonymity on Dark Web Marketplaces. Cryptocurrencies such as Bitcoin provide strong anonymity to their users and

DOI 10.18293/DMSVIVA2023-017

transactions. Thus, it became a vital technology in opening the first Dark Web marketplace; Silk Road [11].

A. The Onion Router (TOR)

The very foundation of the Dark Web and its activities, such as its marketplaces, are based on Onion Routing. The TOR project was created and launched by the US Navy in 2002 to warrant networked anonymous communication [22]. Initially, TOR was created to avoid political censorship and enable freedom of speech over the internet but has since been adapted to facilitate various other activities, including illegal activities [23]. TOR was designed as a low-latency network, a network developed to handle a high capacity of data messages with very little delay or latency while performing functions such as web browsing [24]. The anonymity aspect is achieved through the concept of onion routing. Onion routing allows users to redirect their internet traffic through other users' devices such that the identity of the original user cannot be differentiated from the various other users [25].

Since an essential factor behind anonymity with TOR is to mask one's identity in a sea of various other identities, the soundness of one's anonymity on TOR depends on the number of users on the system [24]. It also depends on whether users on the TOR system are undetectable. If a particular user becomes de-anonymized on the TOR network, this decreases the anonymity level for other users putting the entire network at risk (25). Thus, a knowledgeable understanding of how to download, install and correctly use the TOR software is crucial in ensuring TOR's anonymous integrity [24]. With the demand for online anonymity increasing, TOR has improved its ease of use and provided a mobile version of the software [26]. The Inevitability theory of technology states that once a technology is created, what comes after is its inevitable development [28]. Therefore, TOR's development and advancements will only continue and, if not regulated correctly, could pose challenges for law enforcement.

B. Cryptocurrencies

Cryptocurrencies are a form of digital money that enables users to conduct peer-to-peer (P2P) transactions without the need for centralization [29]. Since cryptocurrencies are a decentralized technology, government and banking institutes have no control. Furthermore, cryptography and blockchain technologies ensure the privacy and security of users and their transactional information [30]. One of the first cryptocurrencies created was Bitcoin in 2009 [7]. As Bitcoin provides a level of security to its users and transactions, and as the cryptocurrency is a decentralized technology, it became a crucial technology in opening the first dark web marketplace, Silk Road. Silk Road's users purchased various illegal items from the marketplace, with payment being made in Bitcoin [11]. Following the fall of Silk Road, various other marketplaces began to rise, and all made use of cryptocurrencies to facilitate their transactions [11]. Since the establishment of cryptocurrencies, there has been a spike in cybercrime worldwide. Having these Dark Web Marketplaces hosted on TOR and facilitated by Bitcoin transactions made it almost impossible for law enforcement and government to regulate illegal activity on marketplaces [30]. A challenging aspect for government and law enforcement regarding cryptocurrencies is that they are decentralized. Meaning no central entity controls it, making it difficult to establish a regulatory framework [32].

C. Decentralization

Most software applications developed adhere to the centralized client-server model by where a central system controls the application. Few applications follow a distributed approach, but very few software applications are decentralized [33]. Decentralized applications can function in two ways, either run on blockchain technology (which is based on peer-to-peer communication) or in a peer-to-peer (P2P) network itself. A significant drawback to Dark Web marketplaces is that they are a centralized entity, meaning a central figure (marketplace admin) controls marketplace activity. A decentralized version of these marketplaces would mean cutting out the middleman and having buyers and sellers interact directly with one another [34].

D. Law Enforcement and Government Intervention

This literature review has highlighted the various technical factors that make the Dark Web and its marketplaces challenging to govern and regulate. Factors such as blockchains and cryptocurrencies decentralized nature where no central entity controls them [32]. Or how TOR's anonymity and encryption attribute inhibit law enforcement in locating cybercriminals [38]. [40] proposes two solutions in decreasing Dark Web-related crime. Solution 1 would be to block access to TOR. Although this will significantly reduce Dark Web-related crime, it would be unfeasible as TOR has a wide range of use cases. The second solution would be to target hidden services. This solution does not have severe repercussions as the first, but it would be more challenging to implement [40].

E. Conceptual Model and Hypotheses

Fig. 1 below presents a conceptual model, of the technological factors influencing the nature of dark marketplaces.

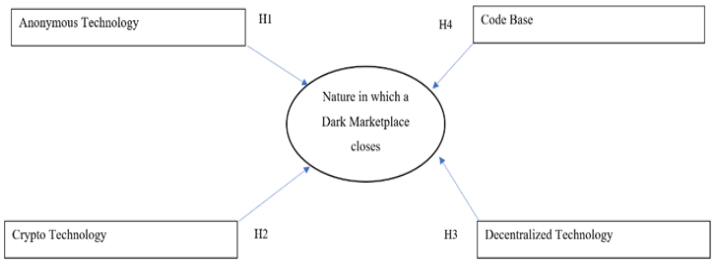


Fig 1. Conceptual model for the study.

The study hypothesises the following:

H1: The type of anonymizing software used to access a dark web marketplace site will influence or determine the nature in which that dark marketplace becomes out of service.

H2: The cryptocurrency used to purchase illegal goods and services from a dark web marketplace will influence or

determine the nature in which that dark marketplace becomes out of service.

H3: There is an association between marketplaces supporting decentralization and the nature in which the marketplace becomes out of service.

H4: The software in which dark web marketplaces are developed will influence or determine the nature in which that dark marketplace becomes out of service.

III. METHODOLOGY

The researchers adopted an objectivism ontological stance and positivism epistemology to guide this study [44], [45]. The study was cross-sectional, and data was collected via a secondary quantitative research method. Collecting large data sets over an extended period of time pertaining to Dark Webs and Dark Web Marketplaces is unfeasible due to time constraints [46]. According to [47], conducting Dark Web data collection requires the researcher to have extensive technical knowledge with web scraping, crawling tools, and routing software. [47] also states that if collecting Dark Web data in such a manner is not feasible, researchers can draw information from digital archives. Material that could be utilized to study Dark Web architecture includes Dark Web forums, mailing lists, hidden sites, and software repositories [47]. Hence classifying the secondary data as multiple-source secondary data as data can be collected from both survey and documentary secondary data [46].

A. Sampling

Obtaining data related to an entire population or all the Dark Web marketplaces functioning on the Dark Web is impractical [42]. The impracticality is derived from the difficulty of identifying all marketplaces on the Dark Web due to the technologies used to keep sites hidden [5]. In such cases, a sampling technique would allow the researcher to only source data on a subset of an entire population or subset of all existing Dark Web marketplaces [50].

As this research deals with the Dark Web and secondary data collection, purpose sampling was adopted [51]. Obtaining data related to the Dark Web can be challenging with all the technical expertise required for collection procedures such as web scraping and crawling, making a purpose sampling strategy suitable. The drawbacks, however, of such a sampling strategy is that it becomes difficult to create a generalization for a population based off the subset chosen.

B. Data Collection

Secondary data was collected from Darknet Market Archives (DNM), an online dark web repository containing Information regarding Dark Web Marketplaces [48]. The DNM archive has been publicly released, and information relating to 87 Darknet Marketplaces was collected [48]. Data relating to the marketplace's technological capabilities, such as the anonymizing software being utilized, the types of cryptocurrencies used to finance transactions being utilized, the codebase in which the marketplace was developed, and whether marketplaces started implementing decentralized technologies were collected. In addition to this, data relating to law enforcement and government efforts were also collected from the DNM, such as the reasoning behind marketplace closing and the success rate of a law enforcement raids.

The data were thus placed into six categorical variables: anonymization, cryptocurrency, decentralisation, codebase, reasoning for marketplace closure, and law enforcement success rate, similar to the original DNM Archive [48]. Each category relates to a section discussed in the literature review chapter of this study where both reasoning for marketplace closure and law enforcement success rate refers to *Law enforcement and Government Intervention.*

To ensure the quality and suitability of the data for analysis, the researchers took several measures to mitigate the potential risk of using inconsistent data. First, reliability and validity tests to evaluate the data's quality and consistency were conducted. While the data size may be limited due to the exploratory nature of the study, the reliability and validity tests suggested fair reliability, indicating that the data could be used for analysis. However, the researchers acknowledged that the data collected from web scrapes and crawls can be prone to external factors such as internet connectivity issues and bugs in the crawling software. Therefore, they took additional steps to verify the accuracy of the data by cross-referencing it with other sources where possible. Overall, these measures helped to ensure the quality and reliability of the data, reducing the risk of confounding the study results. [49].

C. Ethical Considerations

The researchers obtained for ethics approval from the university of Cape Town. Internet-mediated research uses the internet or computing device to conduct archival research and collect secondary data. Secondary data collected and published in a public setting, such as an online data repository, does not require extensive ethical considerations [53]. However according to [46], the sources from which the data was collected should also be distinctly acknowledged, and, if provided, citation guidelines offered by the online repositories should be adhered to.

D. Assessing Reliability and Validity

To assess the reliability and validity of the secondary data collected, it was recommended by [54] to identify copyright statements and published papers utilizing the data. The data set released by [48] was released under the Creative Commons CC0 "No Rights Reserved" license. The Creative Commons (CC) license is a copyright license that defines how information can be distributed. It is utilized in cases where the owner of a piece of work wants to give free access to their work with the intention for their work to be built upon by other users [55]. [54] proclaims that obtaining the data source's copyright statement indicates who is accountable for the data. By obtaining the publications in which the dataset is being utilized, according to [54], will assert the data's reliability as publications are deemed further reliable.

E. Data Analysis

Data analysis was performed using IBM Statistical Package for the Social Sciences (SPSS). A descriptive analysis was first

performed to give a basic description of the data collected. This involved describing each variable's distribution, its central tendency, and the relationship between variables. This analysis is presented in the form of a frequency distribution table, bar charts, and cross-tabulation tables [59]. Following this, regression analysis and hypotheses testing were performed. This allowed for the hypotheses presented in chapter two to be tested by performing a Fisher-Freeman-Halton Exact Test. And for the research question to be answered by performing a linear regression analysis.

IV. RESULTS

A. Descriptive Analysis

Table 1 below presents the characteristics of the data collected from the DNM. Information on 87 Dark Web Marketplaces was collected. The majority of the marketplaces assessed implemented only TOR as their primary anonymizing software (94.3%), with only a small portion of marketplaces utilizing I2P with or instead of TOR (5.7%) as a means of ensuring anonymity on the Dark Web. Bitcoin was the most popular cryptocurrency used to finance transactions, with 89.7% of marketplaces accepting bitcoin as payment for goods and services. Some marketplaces did offer other forms of payment, such as Litecoin (3.4%) and alternative coins, referred to as 'other' (2.3%). However, a select group of marketplaces offered more than one form of payment, allowing their users the option of either paying in Bitcoin or Litecoin (4.6%). More than half of the marketplaces identified did not offer support for multi signatures (80.5%). The codebase in which marketplaces were developed in was unknown for some marketplaces (29.7%). Other marketplaces did make use of open-source PHP frameworks such as Bitwasp (16.1%) and Nette (2.3%) to develop their site. However, a significant portion of marketplaces did decide to custom build their marketplace site (31%). In terms of the reasoning as to why Dark Web Marketplaces stopped operating, 40.2% seized its operations due to scams conducted by marketplace operators, 16.1% were because of hacks that forced closure, 6.9% of marketplaces closed for unknown reasons, and 26.4% were voluntary closures by marketplace operators. From all the 87 marketplaces assessed, only 10.3% were brought to closure by law enforcement. And an even smaller percentage (8%) resulted in the prosecution of marketplace operators.

TABLE I. Frequency distribution describing data collected

Anonymization	**Frequency**	**Percentage**
TOR	82	94.3
I2P with or instead of TOR	5	5.7
Cryptocurrency	**Frequency**	**Percentage**
Bitcoin	78	89.7
Litecoin	3	3.4
Both Bitcoin and Litecoin	4	4.6
Other	2	2.3
Decentralization	**Frequency**	**Percentage**
Made use of Multi signatures	17	19.5
Did not make use of Multi signatures	70	80.5
Codebase	**Frequency**	**Percentage**
Custom	27	31
Bitwasp	14	16.1
Nette	2	2.3
Other	19	21.8
Unknown	25	29.7
Reasoning for Marketplace Closure	**Frequency**	**Percentage**
Law enforcement Raid	9	10.3
Hacked	14	16.1
Scam	35	40.2
Voluntary	23	26.4
Unknown	6	6.9
Law Enforcement Success Rate	**Frequency**	**Percentage**
Led to prosecution	7	8.0
Did not lead to prosecution	80	92.0

As is the case for descriptive data, to measure the central tendency, which is to identify the most frequent value, the mode, will be most appropriate [59]. Table 2 below presents the mode for each category in the data set. For the anonymization category, TOR was the most frequently used anonymizing tool. Bitcoin was identified as the most popular cryptocurrency to purchase goods and services from marketplaces. Custom-built marketplaces were the most common choice taken by marketplace operators when developing the marketplace site. Scams were the usual way in which marketplaces closed. And finally, out of all the 87 illegal marketplaces, 80 of them did not face any legal repercussions.

TABLE II. The Mode for each categorical variable in the dataset

Category	**Mode (Count out of 87)**
Anonymization	TOR (82)
Cryptocurrency	Bitcoin (78)
Decentralization	Did not make use of multi signatures (70)
Codebase	Custom (27)
Reasoning for Marketplace Closure	Scam (35)
Law Enforcement Success Rate	Did not lead to prosecution (80)

B. Interdependence between Anonymizing Technology and Marketplace Closure Reasoning

A cross-tabulation analysis will be suitable to analyse the interdependence between two variables [46]. Table 3 below

illustrates the interdependence between the anonymizing technology used and the nature in which the marketplace closed. Overall, 10.3% of marketplaces were using some form of anonymizing software and getting raided by law enforcement. However, 11% of marketplaces implementing TOR succumbed to a law enforcement raid which is more than the total amount of marketplaces getting raided (10.3%). An adjusted residual value above 2 indicates that the observed frequency for a particular cell is more than the frequency expected for that cell. An adjusted residual value below -2 suggests that the observed frequency for a specific cell is smaller than the frequency expected for that cell [60]. As the adjusted residual is either below 2 or above -2, there is no deviation explaining that 11% is not statistically differentiable from the total of 10.3%. This interpretation was similar for reasoning consisting of Hacked, Scam, Voluntary and Unknown, where the percentage within Anonymizing Technologies is not statistically differentiable from the total value as the adjusted residual values are either below 2 or above -2.

TABLE III. Anonymizing Technology and Marketplace Closure Reasoning

Reason for Marketplace Closure	Anonymizing Technologies		Total
	TOR	I2P with or instead of TOR	
Raid			
Count (%)	9 (11%)	0 (0.0%)	9(10.3%)
Adjusted Residual	0.8	-0.8	
Hacked			
Count (%)	14 (17.1%)	0 (0.0%)	14(16.1%)
Adjusted Residual	1.0	-1.0	
Scam			
Count (%)	34 (41.5%)	1(20.0%)	35(40.2%)
Adjusted Residual	1.0	-1.0	
Voluntary			
Count (%)	20 (24.4%)	3(60.0%)	23(26.4%)
Adjusted Residual	-1.8	1.8	
Unknown			
Count (%)	5(6.1%)	1(20.0%)	6(6.9%)
Adjusted Residual	-1.2	1.2	
Total			
Count (%)	82(100%)	5(100%)	87(100%)

C. *Interdependence between Crypto Technology and Reasoning for Marketplace closure*

Table 4 below illustrates the interdependence between the crypto technology used to finance transactions and the nature in which the marketplace closed. It is evident that the percentage within Cryptocurrencies for Raid, Hacked, Scam, and Voluntary is not statistically differentiable from the total value as the adjusted residual values are either below 2 or above -2. However, 50% of marketplaces that supported both Bitcoin and Litecoin closed for unknown reasons, more than the total amount of marketplaces closing for unknown reasons (6.9%). As the adjusted value is greater than 2 (3.5), significantly more marketplaces support both Bitcoin and Litecoin than expected if there was no dependency between variables.

TABLE IV. Crypto Technology and Reasoning for Marketplace closure

Reason for Marketplace Closure	Cryptocurrencies				Total
	Bitcoin	Litecoin	Bitcoin and Litecoin	Other	
Raid					
Count (%)	8 (10.3%)	1(33.3%)	0(0.0%)	0(0.05%)	9(10.3%)
Adjusted Residual	-.1	1.3	-.7	-.5	
Hacked					
Count (%)	14(17.9%)	0(0.0%)	0(0.0%)	0(0.0%)	14(16.1%)
Adjusted Residual	1.4	-.8	-.9	-.6	
Scam					
Count (%)	31(39.7%)	1(33.3%)	1(25.0%)	2(100.0%)	35(40.2%)
Adjusted Residual	-.3	-.2	-.6	1.7	
Voluntary					
Count (%)	21(29.5%)	1(33.3%)	1(25.0%)	0(0.0%)	23(26.4%)
Adjusted Residual	.3	.3	-.1	-.9	
Unknown					
Count (%)	4(5.1%)	0(0.0%)	2(50.0%)	0(0.0%)	6(6.9%)
Adjusted Residual	-1.9	-.5	3.5	-.4	
Total					
Count (%)	78(100%)	3(100%)	4(100%)	2(100%)	87(100%)

D. *Interdependence between Decentralized Technology and Reasoning for Marketplace closure*

Table 5 below illustrates the interdependence between marketplaces implementing decentralized technologies by offering support for multi signatures and the nature in which the marketplace closed.

TABLE V. Decentralized Technology and Reasoning for Marketplace closure

Reason for Marketplace Closure	Decentralized Technologies		Total
	Did not support Multiple Signatures	Support Multiple Signatures	
Raid Count (%) Adjusted Residual	6(8.6%) -1.1	3(17.6%) 1.1	9(10.3)
Hacked Count (%) Adjusted Residual	12(17.1%) .5	2(11.8%) -.5	14(16.1%)
Scam Count (%) Adjusted Residual	28(40.0%) -.1	7(41.2%) .1	35(40.2%)
Voluntary Count (%) Adjusted Residual	19(27.1%) .3	4(23.5%) -.3	23(26.4%)
Unknown Count (%) Adjusted Residual	5(7.1%) .2	1(5.9%) -.2	6(6.9%)
Total Count (%) Adjusted Residual	70(100%)	17(100%)	87(100%)

E. Regression Analysis

To perform a regression analysis with categorical input variables, each variable will subsequently be transformed into dummy variables. This involves coding the data as 1's and 0s. Where 1 refers to a data point that belongs to a category and 0 for all data points that do not belong. Thus, treating the categorical input variables as a continuous variable for analysis [61]. Presented in table 6 are the results of the regression analysis performed for each independent variable on the dependant variable, represented as two values r (coefficient of correlation) and r^2 (coefficient of determination). The R-value for each variable ranges between 0 and 0.2, indicating a weak but positive correlation between the variables and a low likelihood for the dependant variable to be influenced by the independent variable. With crypto technology having the most significant influence on the nature in which a Dark Marketplace closes. The r^2 values indicate that 0.7% of the dependant variable is predicted by anonymous technology, 2.8% is predicted by crypto technology, 3.7% is predicted by the codebase in which marketplaces are developed, and 1.4% by decentralized technology.

TABLE VI. Results of regression analysis

Independent Variable	Dependent Variable	
	Nature in which a Dark Marketplace closes	
	r	r^2
Anonymous Technology	0.084	0.007
Crypto Technology	0.166	0.028
Code Base	0.091	0.037
Decentralized Technology	0.118	0.014

F. Hypotheses Testing

A Fischer's Exact Test of Independence was deemed appropriate to test the hypotheses (62). A Fischer's Exact Test of Independence is also recommended for analysis in situations where cross-tabulation tables are of 2x2 matrices, and the sample size of the data set is less than 1000, in this case, 87 (63).

Presented in table 7 are the results of conducting the Fischer's Exact test on each of the hypotheses. A probability value P of less than 0.05 indicates a significant association between the independent and dependant variables (46). Based on the Fischer's Exact test conducted in SPSS and presented in table 7, it is evident that none of the four hypotheses established have a significant association between the independent and dependant variables, indicating that neither of the four hypotheses was supported. H1 (P-value= 0.221), H2 (P-value = 0.277), H3 (P-value = 0.859), and H4 (P-value = 0.828) all have p-values greater than 0.05.

TABLE VII. Results of Hypotheses Testing

Hypotheses	Independent Variable	Dependent Variable	Fisher-Freeman-Halton Exact Test		Supported ?
			Value	P-value	
H1	Anonymous Technology	Nature in which Dark Marketplace closes	4.655	0.221	No
H2	Crypto Technology	Nature in which Dark Marketplace closes	11.941	0.277	No
H3	Decentralized Technology	Nature in which Dark Marketplace closes	1.595	0.859	No

H4	Code Base	Nature in which Dark Marketplace closes	11.208	0.828	No

V. DISCUSSION OF FINDINGS

Anonymizing technologies serve as the foundation to the dark web and its marketplaces, with TOR being the most popular tool identified within literature [17]. TOR's popularity was justified during analysis, with a large majority of the 87 marketplaces sampled using TOR and only a select few marketplaces implementing I2P. This can be attributed to the reliable anonymity that TOR provides and its constant improvement, with developments made to its ease of use and enabling mobile access to the software [26]. Marketplaces closing because of scams were most prevalent for TOR, whereas marketplaces utilizing I2P were mostly voluntary closures from the marketplaces sampled. However, H1 results show that there was no significant evidence in determining whether the type of anonymizing technology being used within a marketplace will influence the way in which that marketplace closes. Thus, as [32] and [64] discussed, establishing an overarching regulatory framework will be appropriate as this can target the use cases of such technologies. Summarily for crypto technology, with Bitcoin being the most popular among marketplaces in financing anonymous transactions and some marketplaces also incorporating Litecoin to mitigate the slow clearance rate of transactions. Yet, according to the results of H2, there was no significant association between the type of cryptocurrency utilized and the nature in which the marketplace closed. Again, this illustrates that instead of regulating a specific cryptocurrency, policies relating to digital currencies and their use cases should be developed [32].

Dark web marketplaces are vulnerable to closures due to their centralized nature, with a central entity managing the marketplace [65]. Decentralized applications, on the other hand, are more resistant to closures. OpenBazaar, for example, uses multi signatures to enable decentralization, but most other marketplaces analyzed do not support multi signatures. This is because many marketplaces follow a centralized client-server model. This was consistent in literature as many applications adhere to a centralized client-server model [33]. Hence, there was no association between marketplaces supporting decentralization and the nature in which the marketplace had closed thus supporting the results of H3.

The software used in the development of marketplace can mitigate the risk of closures. Java-based codebases are a good solution, as they allow marketplaces to migrate from server to server as many marketplaces analyzed adhered to a centralized client-server model [66]. However, several marketplaces use PHP Frameworks such as Nette and Bitwasp. While Bitwasp supports multi signatures and allows marketplaces developed with the framework to function independently of central servers, H4 results showed that the type of codebase used by marketplaces does not necessarily determine their ability to resist closures or support decentralization.

Each technological factor, anonymization, cryptocurrencies, decentralization, and codebase had a low likelihood of influencing how a marketplace seized its operations, with crypto technology having the greatest significance out of the four technologies identified. This is not unexpected as cryptocurrencies provide both strong anonymity and the ability to enable the sale of illegal goods on the dark web.

VI. CONCLUSION

The study conducted aimed at understanding how dark web marketplaces operate, especially in terms of the various technologies and their impacts on marketplace closures. It was identified in the literature that anonymization provided with software such as TOR and cryptocurrencies like Bitcoin are fundamental components in enabling marketplace activity. This was echoed in the findings of this study as both TOR and Bitcoin were extensively applied throughout marketplaces. The codebases and decentralization were then characterized as additional techniques to mitigate against shortcomings such as the single point of failure with centralized applications. However, the findings of the study confirmed that despite centralization being such a pitfall for marketplaces, most marketplaces still opted not to implement multi signatures or develop marketplaces with codebases that supported it. The technological factors linked to dark web marketplaces closures all had a low probability of determining how a marketplace would become out of service. However, crypto technology was found to have the most impact in allowing dark web marketplaces to operate, thus, illustrating the importance of effective regulation of crypto technology, focusing on its use cases to reduce illegal online activity.

VII. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

The secondary data collected to conduct this research study was bounded by the period in which the data was initially collected. According to the Darknet Market Archives (DNM) dataset, its last known update date was June the 9th, 2019. Many archival datasets relating to the dark web are not updated regularly because of the difficulty and cost of web scraping and crawling dark websites. And the potential for newer techniques or technologies to be implemented within marketplaces has not been accounted for. As technological advancements are rapidly increasing, future research should consider the timeframe in which data collection took place. If possible, primary data should be collected and made available to the public to extend the research opportunities to fields that do not possess the required technical expertise to collect the data.

REFERENCES

[1] D. S. Rudesill, J. Caverlee, and D. Sui, "The deep web and the Darknet: A look inside the internet's massive black box," Woodrow Wilson International Center for Scholars, STIP, vol. 3, pp. 1-17, October 2015.

[2] D. Kolb, "Surface Web is Only the Tip of the Iceberg," Traversals, Available: https://traversals.com/blog/surface-web/. [Accessed: June. 23, 2022].

[3] J. Frankenfield, "*Deep Web*." Investopedia. Available: https://www.investopedia.com/terms/d/deep-web.asp. [Accessed: December. 24, 2022].

[4] Kaspersky, "What is the Deep and Dark Web?" Available: https://www.kaspersky.com/resource-center/threats/deep-web. [Accessed: December. 14, 2021].

[5] E. Jardine, "The Dark Web dilemma: Tor, anonymity and online policing. Global Commission on Internet Governance Paper Series, No. 21, pp. 1-13, September 2015.

[6] A. Gupta, S. Maynard, and A. Ahmad, "The Dark Web as a Phenomenon: A Review and Research Agenda," in the Proc. 30th Australasian Conference on Information Systems, Perth, Australia, pp. 1-12, December 2019.

[7] C. Dipiero, "Deciphering cryptocurrency: Shining a light on the deep dark web," University of Illinois law review. vol. 3, pp. 1267-1299, March 2017.

[8] S. Sadik, and M. Ahmed, "An overview of the Dark Web," in *Security Analytics for the Internet of Everything,*, M. Ahmed, U. Barkat, and A. Pathan, Eds. New York: CRC Press, 2020, pp. 55-66.

[9] Z. Mador, "Keep the dark web close and your cyber security tighter," Comput. Fraud Secur, no. 1, pp. 6–8, January 2021

[10] K. Soska, A. Kwon, A. Christin, N. Devadas, and S. Beaver, "A decentralized anonymous marketplace with secure reputation," Technical Report 2016/464, IACR Cryptology ePrint Archive, pp. 1-15, 2016.

[11] D. Stroukal, and B. Nedvedová, "Bitcoin and other cryptocurrency as an instrument of crime in cyberspace," Proc. of 4th Business and Management Conferences, Istanbul, vol. 4407036, pp. 219-226, October 2016.

[12] S. He, Y. He, and M. Li, "Classification of Illegal Activities on the Dark Web," Proceedings of the 2019 2nd International Conference on Information Science and Systems, Taiyuan, China, pp. 73-78, March 2019.

[13] I. Ladegaard, "Open Secrecy: How Police Crackdowns and Creative Problem-Solving Brought Illegal Markets out of the Shadows," Soc Forces, vol. 99, pp 532-559, November 2020.

[14] Dnstats.net. "List of Darknet Markets in 2021." Available: https://dnstats.net/list-of-darknet-markets/[Accessed: December. 24, 2022].

[15] N. House, "The 2021 Guide to Darknet Markets. The Cyber Security Company. Available: https://www.stationx.net/the-2021-guide-to-darknet-markets/, Jan. 14, 2022 [Accessed: December 15, 2022]

[16] E. Jardine, "Tor, what is it good for? Political repression and the use of online anonymity-granting technologies," New media Soc, vol. 20, pp. 435-452, February 2018.

[17] S. Winkler, and S. Zeadally, "An analysis of tools for online anonymity," Int. J. Pervasive Comput. Commun., vol. 11, pp. 436-453, November 2015.

[18] R. Kang, S. Brown, and S. Kiesler, "Why do people seek anonymity on the internet?," Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France, pp. 2657-2666, April 2013.

[19] S. Larsson, M. Svensson, M. de Kaminski, K. Rönkkö, and J. Alkan Olsson, "Law, norms, piracy and online anonymity," J. Interact. Mark., vol. 6, pp. 260–280, October 2012.

[20] H. Arora, "Possible Silver Lining for the Content-Owners in Illegal File-Sharing Acts," Available at SSRN 3614389, May 2020.

[21] A. D. Berkowitz, "Applications of social norms theory to other health and social justice issues," in The social norms approach to preventing school and college age substance abuse: A handbook for educators, counselors, and clinicians, H. W. Perkins, Ed. Jossey-Bass:Wiley, 2003, pp. 259–279.

[22] A. S. Beshiri and A. Susuri, "Dark Web and Its Impact in Online Anonymity and Privacy: A Critical Analysis and Review," JCMC, vol. 07, pp. 30–43, 2019.

[23] A. Chaabane, P. Manils, and M. A. Kaafar, "Digging into anonymous traffic: A deep analysis of the tor anonymizing network," in Proceedings of the 2010 4th International Conference on Network and System Security. IEEE Computer Society, Washington, DC.

[24] K. Gallagher, S. Patil, and N. Memon, "New me: Understanding expert and non-expert perceptions and usage of the Tor anonymity network.," 13th Symposium on Usable Privacy and Security, Santa Clara, California, SOUPS 2017,

[25] J. Clark, P. C. Van Oorschot, and C. Adams, "Usability of anonymous web browsing: an examination of tor interfaces and deployability," Conference Proceedings of the 3rd Symposium on Usable Privacy and Security, Pittsburgh, Pennsylvania, USA SOUPS 2007.

[26] A. Nastuła, "Dilemmas related to the functioning and growth of Darknet and the Onion Router network.," Journal of Scientific Papers "Social development and Security," vol. 10, pp. 3–10, April 2020.

[27] G. N. Nedeltcheva, E. Vila, and M. Marinova, "The Onion Router: Is the Onion Network Suitable for Cloud Technologies'" in Smart Technologies and Innovation for a Sustainable Future: Advances in Science, Technology & Innovation, A. Al-Masri and K. Curran, Eds. Springer: Cham., 2019.

[28] M. Cuneta "Bitcoin's Inevitability Thesis, Understanding the unstoppable nature of technology". Available: https://medium.com/@MiguelCuneta_21450/bitcoins-inevitability-thesis-d89585e62356, April, 2019, [Accessed: June 14, 2021].

[29] S. Lee, C.Yoon, H. Kang, Y. Kim, Y. Kim, D. Han, S. Son, and S. Shin, "Cybercriminal Minds: An investigative study of cryptocurrency abuses in the Dark Web," Proceedings of 2019 Network and Distributed System Security Symposium, 24-27 San Diego, CA, USA, February 2019.

[30] M. Milutinović, "Ekonomika," Journal for Economic Theory and Practice and Social Issues, vol. 64, pp 105-122, 2018.

[31] A. Barysevich, & A. Solad, "Litecoin emerges as the next dominant dark web currency. Recorded Future." Available: https://www.recordedfuture.com/dark-web-currency/, March 8, 2018, [Accessed: November 12, 2022].

[32] H. Nabilou, "How to regulate bitcoin? Decentralized regulation for a decentralized cryptocurrency," Int. J. Law Inf. Technol., vol. 27, pp. 266–291, 2019.

[33] S. Raval, "Decentralized applications: harnessing Bitcoin's blockchain technology." O'Reilly Media, Inc. 2016.

[34] Hussey, M, 'What are decentralized marketplaces?" Available: https://decrypt.co/resources/what-are-decentralized-marketplaces, March, 2020, [Accessed: December 15, 2022]

[35] A. Greenberg, "Inside the 'DarkMarket' Prototype, a Silk Road the FBI Can Never Seize." Wired. Available: https://www.wired.com/2014/04/darkmarket/, April 24, 2014, [Accessed: December 10, 2022].

[36] I. Allison, "Mover over eBay: Countdown to OpenBazaar and the decentralised marketplace revolution." International Business Times. Available: https://www.ibtimes.co.uk/move-over-ebay-countdown-openbazaar-decentralised-marketplace-revolution-1529767, November 20, 2015, [Accessed: December 24, 2022].

[37] J. Redman, "Meet Beaver: A Decentralized Anonymous Marketplace." Bitcoin News, Available: https://www.livebitcoinnews.com/meet-beaver-a-decentralized-

anonymous-marketplace/, May 19, 2016, [Accessed: December 18, 2022].
[38] R. Heaton, "How does Tor work?" Available: https://robertheaton.com/2019/04/06/how-does-tor-work/, April 6, 2019, [Accessed: December 18, 2022].
[39] A. Ghappour, "Searching Places Unknown: Law Enforcement Jurisdiction on the Dark Web," Stanford Law Review, vol. 69, April 2017.
[40] N. V. Denic, "Government Activities to Detect, Deter and Disrupt Threats Enumerating from the Dark Web," Technical Report, US Army Command and General Staff College Fort Leavenworth United States, 2017.
[41] J. Dalins, C. Wilson, and M. Carman, "Criminal motivation on the dark web: A categorisation model for law enforcement," Digit Investig, vol. 24, pp. 62–71, March 2018.
[42] R. V. Clarke, "Situational Crime Prevention," Crime and Justice, vol. 19, pp. 91–150, January 1995.
[43] K. Hegadekatti, "Regulating the Deep Web Through Controlled BlockChains and Crypto-Currency Networks," SSRN Electronic Journal, pp. 1-10, December 2016.
[44] A. Ahmed, "Ontological, Epistemological and Methodological Assumptions: Qualitative versus Quantitative," Online Submission, pp. 1-13, 2008.
[45] H. Collins, "Creative research: the theory and practice of research for the creative industries," Bloomsbury Publishing: New York, pp. 1-203, 2018.
[46] M. Saunders, P. Lewis, and A. Thornhill, Research *Methods* for *Business Students*. Pearson: New York, 2012.
[47] R. W. Gehl, "Archives for the Dark Web: A Field Guide for Study," in Research Methods for the Digital Humanities, L. Levenberg, T. Neilson and D. Rheams, Eds. Springer Nature: Switzerland AG, 2018, pp. 31–51.
[48] G. Branwen, N. Christin, D. Decary-Hetu, Munksgaard R. Andersen, E. Presidente, Anonymous, Lau, D., Sohhlz, Kratunov, D., Cakic, V., A. Buskirk, Whom, M. Mckenna, & Goode, "Dark Net Market archives, 2011-2015." Available: https://www.gwern.net/DNM-archives S., March 20, 2021, [Accessed: November 20, 2022].
[49] M. P. Johnston, "Secondary data analysis: A method of which the time has come." Qualitative and quantitative methods in libraries, vol. 3, pp. 619-626, September 2014.
[50] M. Saunders, P. Lewis, and A. Thornhill. "Research *Methods* for *Business Students*. Pearson: New York, 2009.
[51] I. Etikan, "Sampling and Sampling Methods," BBIJ, vol. 5, pp. 210-213, May 2017.
[52] A. S. Acharya, A. Prakash, P. Saxena, and A. Nigam, "Sampling: why and how of it?," Indian Journal of Medical Specialities, vol. 4, pp 330-333, July 2013.
[53] L. Cilliers, and K. Viljoen, "A framework of ethical issues to consider when conducting internet-based research," SAJIM, vol. 23, pp. 1-9, March 2021.
[54] N. Ó. Dochartaigh, "Internet Research Skills (3rd ed.)," SAGE Publications, Inc., 2012.
[55] G. Hagedorn et al., "Creative Commons licenses and the non-commercial condition: Implications for the re-use of biodiversity information," ZooKeys, vol. 150, pp. 127–149, November 2011.
[56] K. Kruithof, J. Aldridge, D. Hétu, M. Sim, E. Dujso, and S. Hoorens, "Internet-facilitated drugs trade: An analysis of the size, scope and the role of the Netherlands," Rand Corporation: Cambridge, UK, 2016.
[57] S. Ghosh, A. Das, P. Porras, V. Yegneswaran, and A. Gehani, "Automated categorization of onion sites for analyzing the darkweb ecosystem," Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS: Canada 2017.
[58] P. H. Meland, Y. F. F. Bayoumy, and G. Sindre, "The Ransomware-as-a-Service economy within the darknet," Computers & Security, vol. 92, p. 101762, May 2020.
[59] M. K. William, "Research Methods Knowledge Base: Descriptive statistics." Available: https://conjointly.com/kb/descriptive-statistics/, 2021, [Accessed: November 12, 2022].
[60] A. Agresti, "Categorical Data Analysis," Wiley Series in Probability and Statistics, New York: Wiley, 2002.
[61] H. Schielzeth, "Simple means to improve the interpretability of regression coefficients," Methods Ecol. Evol., vol.1, pp. 103–113, February 2010
[62] G. M. Gaddis, and M. L. Gaddis, "Introduction to biostatistics: Part 5, statistical inference techniques for hypothesis testing with nonparametric data," Ann Emerg Med, vol. 19, pp. 1054–1059, September 1990.
[63] L. M. Connelly, "Fisher's exact test," Medsurg Nursing, vol. 25, pp. 58-60, 2016.
[64] A. Spithoven, "Theory and Reality of Cryptocurrency Governance," J. Econ. Issues, vol. 53, pp. 385–393, April 2019.
[65] L. Brittney, "Deep Dot Web Seized. Terbium Labs." Available: https://terbiumlabs.com/2019/07/11/the-king-is-dead-long-live-decentralized-markets/ 2019, [Accessed: November 12, 2022].
[66] M. Shoaib, A. Ishaq, M. Awais, S. Talib, G. Mustafa, and A. Ahmed, "Software Migration Frameworks for Software System Solutions: A Systematic Literature Review," International Journal of Advanced Computer Science and Applications, vol. 8, pp. 192-204, 2017.

".