Deep Learning based Information Extraction Framework on Chinese Electronic Health Records

Bing Tian ☆ Yong Zhang ☆ Kaixin Liu ☆ Chunxiao Xing ☆

☆ RIIT, Beijing National Research Center for Information Science and Technology, Department of Computer Science and Technology, Institute of Internet Industry, Tsinghua University, Beijing, China

Abstract

Electronic Health Records (EHRs) store a large amount of clinical data associated with each patient. Information extraction on unstructured clinical notes in EHRs is important which could contribute to huge improvement in patient health management. Previous studies mainly focused on English corpus. However, at the same time there are very limited research work on Chinese EHRs. Due to the challenges brought by the characteristics of Chinese, it is difficult to apply existing techniques for English on Chinese corpus. In this paper, we propose a deep learning based framework for information extraction from clinical notes in Chinese EHRs. Our framework consists of three components: data preprocessing, feature generation and entity and relation extractor. For clinical entity recognition, we propose a novel Conditional Random Field (CRF) based model and introduce effective features by leveraging the characteristics of Chinese language. For relation extraction, we utilize Convolutional Neural Network (CNN) to obtain high quality entity-relation facts. To the best of our knowledge, this is the first framework to apply deep learning to information extraction from clinical notes in Chinese EHRs. We conduct extensive sets of experiments on real-world datasets from hospital. The experimental results show the effectiveness of our framework, indicating its practical application value.

Key words: *Electronic Health Records; Deep learning; Information extraction; Entity recognition; Chinese*

1 Introduction

Electronic Health Records (EHRs) store a large amount of clinical data associated with each patient encounter, including demographic information, current and past diagnoses, prescriptions etc [1]. Information extraction from unstructured clinical notes in EHRs, which serves as the first step towards constructing medical-domain specific knowledge graph, can be beneficial for many fields such as disease inference, clinical decision support systems and risk prediction etc [2, 3, 4]. As such, recently years have seen lots of studies concentrated on information extraction from English clinical notes.

However, when it comes to Chinese domain, very limited work has been done especially for relation extraction due to the challenges brought by the Chinese clinical notes. On one hand, the different characteristics of Chinese language determine that the methods on English corpus can not be directly applied on Chinese documents. For example, there is no blank space representing word boundaries between Chinese words, and words have no morphological changes in different situations. Besides, some Chinese function words which are important for semantic understanding, such as "的", "了" are often omitted. On the other hand, since there are a large number of professional terms, abbreviations and medical-domain based knowledge contained in clinical notes. It is difficult to adopt existing Chinese-based work focusing on other domains, such as Chinese social media [5, 6], to our problem.

To address these challenges, we propose a deep learning based information extraction framework on clinical notes in Chinese EHRs. Our framework contains three major components: data preprocessing, feature generation and entity and relation extractor. For data preprocessing, we clean the raw corpus and invite medical experts to make necessary annotations. For feature generation, we then select high quality features from multiple aspects according to the characteristics of clinical notes and Chinese language. Finally, we adopt such features in a novel CRF-based model to identify boundaries and type of clinical entities in clinical notes. Next we consider the superior performance obtained by deep learning based methods in information extraction these years and creatively utilize the convolutional neural network (CNN) model in our task. CNN has been used in many fields [7]. Compared with state-of-the-art methods on English documents which heavily depend on manual feature engineering [8, 9], our CNN-based model can achieve better performance while avoiding intensive human labor. We conduct extensive experiments on a real world EHR dataset from a famous medical institute. And the ex-

DOI reference number: 10.18293/SEKE2018-040

perimental results demonstrate the effectiveness of our proposed framework.

The rest of paper is organized as follows. Section 2 provides an overview of the existing information extraction approaches. Section 3 introduces our deep learning based information extraction framework. Section 4 and Section 5 respectively describe our clinical entity recognition and relation extraction methods in detail. Section 6 reports the experiments and discusses the results. Finally, we draw our conclusions in Section 7.

2 Related Work

In this section, we first review the related work about information extraction on English clinical notes and then introduce the information extraction methods on Chinese and their applications in health-related domain.

Recently, a large amount of work has focused on information extraction on English clinical notes. Due to the unstructured nature, most work utilize the statistical machine learning methods. For example, Seol et al. [10] proposed a clinical Problem-Action relation extraction framework based on CRF and Support Vector Machine(SVM). Skeppstedt et al. [11] studied the usefulness of features extracted from unsupervised methods and applied them in clinical named entity recognition problem. It is noteworthy that these methods have depended on manually engineered features which have seen limited adoption. As such, some recent studies have proposed several methods using deep learning. Jagannatha et al. [12] regarded clinical named entity recognition as a sequence labeling problem and utilized Recurrent Neural Network(RNN) based model. Sahu et al. [13] focused on extracting relations from clinical discharge summaries and exploited the power of CNN to learn features automatically.

Despite the great challenges of information extraction on Chinese documents, there has been a lot of work focused on it recently. For example, in Chinese social media domain, Peng et al. [5] jointly trained word segmentation with an LSTM-CRF model for named entity recognition problem. He et al. [6] further improved the performance for named entity recognition on the same datasets by proposing a unified model combining cross-domain learning and semi-supervised learning. In health-related domain, Yao et al. [14] focused on the text classification on traditional Chinese medicine(TCM) clinical records and proposed a novel method combining deep learning text representation with TCM domain knowledge. He et al. [15] studied the corpus construction of Chinese clinical texts.

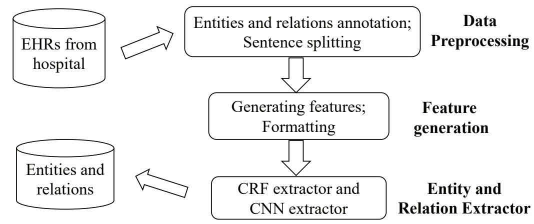

Figure 1: Framework Architecture

患\o者\o 无\o明\o显\o诱\o因\o 出\o现\o胸\ B-symptom 痛\I-symptom，\o服\o硝\B- Treatment 酸\I- Treatment 甘\I- Treatment 油\I- Treatment 可\o缓\o解\o 。\o

Figure 2: BIO Format of Entity Annotation

3 Framework Architecture

The key idea of our framework is extracting clinical entities and relations between them. As shown in Figure 1, there are three main components in our framework: data preprocessing, feature generation and entity and relation extractor. The data preprocessing component processes the raw clinical notes from hospital. Firstly, to generate a high quality corpus for training and testing, we have invited professionals from hospital to help annotate the corpus. And to apply CRF based algorithms to the entity recognition problem, annotated entities should be typically converted into a BIO format. Specifically, it assigns each word into a class as follows: B means the beginning of an entity, I means inside an entity, and O means outside of an entity. For the sentence " 患者无明显诱因出现胸痛，服用硝酸甘油可缓解 "(No obvious cause of chest pain, taking nitroglycerin can relieve symptoms), the BIO format of annotation is shown in Figure 2. Annotated relations are expressed in a triple format [h, r, t], the triple means there exists a relation named r between the entities named h and t. Secondly, considering most relations are existed within one sentence, the preprocessing component splits the clinical notes into sentences using natural language processing tools.

Feature generation component is mainly designed to generate features needed in entity and relation extractor component and normalizes the format of training data so that it can meet the requirements of extractor component. Generally speaking, the data should consist of multiple tokens, and a token consists of multiple columns representing the features.

The entity and relation extractor component learns two extractors: CRF-based clinical entity recognition and CNN-based relation extraction. The two extractor enable extracting clinical entities and relations from clinical notes automatically. Clinical entities and relationships are actually the knowledge contained in clinical notes in health domain so that can be further used in the construction and application of medical-domain specific knowledge graph.

4 Clinical Entity Recognition

In this section, we apply the CRF-based model to Chinese Entity recognition problem. First we introduce the features we choose and then we propose our CRF-based model based on these features.

4.1 Features

According to the characteristics of clinical notes and Chinese language, we select the bag-of-characters feature, Part of Speech(POS) tag feature, and dictionary feature etc. as our feature sets for clinical entity recognition problem.

Bag-of-characters feature As the basic units of Chinese, both characters and phrases can express basic information of Chinese documents. For clinical entity problem, the operations on phrases to generate bag-of-words tend to be more synonymous to complex model than to better performance. So in this paper, we select the bag-of-characters as our feature rather than bag-of-words.

POS tag feature Besides bag-of-characters feature, the POS tag information can help improve the efficiency and precision of clinical entity recognition. Through the analysis of clinical notes, we find that different kinds of clinical entities show different characteristics in the POS tag composition. In addition, usually there will be a verb in front of the entity "test" and "treatment" etc. POS tag features can be generated through the existing natural language processing tools.

Dictionary feature Clinical notes are highly specialized medical relevant texts which contain a large number of medical terminology. Therefore, the introduction of medical entity dictionaries can effectively improve the accuracy of clinical entity recognition. But there are no such dictionaries available in Chinese domain yet. Considering this situation, we construct a Chinese-based medical dictionary as our feature by co-operating with the professionals from hospital. we first extract numerous clinical entities by referring to large amounts of books and literatures as our basic dictionary and then expand it by crawling and filtering data from Internet. The details of the dictionary are shown in Table 1.

4.2 CRF-based Model

In natural language processing domain, CRF is mainly used to solve sequence annotation problems.

Table 1: Medical Entity dictionary

Entity	Example	Number
Disease	胸椎键盘突出 (Thoracic keyboard protrusion)	31450
Medicine	接骨续筋片 (Fracture tablets)	38726
Treatment	齿槽再造术 (Guttural reconstruction surgery)	8493
Test	CT 造影增强扫描 (CT)	3473
Organ	胸口 (Chest)	6089
Physical indicator	血清触珠蛋白 (Serum haptoglobin)	3314

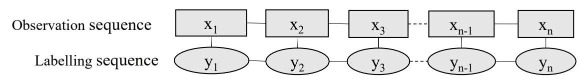

Figure 3: Chain Structure of CRF

Not only can it capture a large amount of human observational experience but also enable capture Markov-chain dependencies between different tags. What's more, by adding customized features according to specific task, CRF has achieved good results on many entity recognition problems.

In this paper, we regard the clinical entity recognition as a sequence labelling problem. Under this situation, we believe that the CRF dependency graph is a chain structure. And what we attempt to do is modelling the conditional probability of multiple variables by giving their observation values. Specifically, as shown in Figure 3, assuming that the observation sequence is $\vec{X} = (x_1,\ x_2,\ \dots\ ,\ x_n)$. $\vec{Y} = (y_1,\ y_2,\ \dots\ ,\ y_n)$ is the corresponding labelling sequence, and y_i means the label of the ith instance of sequence X. Our goal is to construct the conditional probability model $P(\vec{Y} \mid \vec{X})$. Here, $\vec{X}$ is the entire Chinese character sequence of a sentence in the clinical notes. $\vec{Y}$ is the sequence of entity labels corresponding to each word in the sequence $\vec{X}$. And we define our feature function as $f_{a,i}(y_{i-1}, y_i, \vec{X}, i)$. In this function, $a \in A$ represents the type of feature, x_i is the word that we are going to label. λ_a are the corresponding parameters we need to train. For observation sequence $\vec{X}$ and labelling sequence $\vec{Y}$, the conditional probability is as follows:

$$p(\vec{Y}|\vec{X};\lambda) = \frac{1}{Z(\vec{X},\lambda)} \exp(\sum_{a \in A, y_1, y_2 \in \vec{Y}} \lambda_{a,y_1,y_2} \sum_{i=1}^{n} f_{a,i,y_1,y_2}(y_{i-1}, y_i, \vec{X}, i) \tag{1}$$

$Z(\vec{X},\lambda)$ is the regularization term. And the final labelling sequence we get based on this model is as follows:

$$\vec{Y^*} \overset{def}{=} argmax_{\vec{y} \in \vec{Y^n}} p(\vec{y}|\vec{x};\lambda) \tag{2}$$

Table 2: The relations and their occurrence frequencies

Relation	Type	Number
Treatment and disease	治疗施加于疾病 (TrAD)	1460
	治疗改善疾病 (TrID)	260
Treatment and symptom	治疗改善症状 (TrIS)	910
	治疗导致症状 (TrCS)	70
	治疗施加于症状 (TrAS)	2760
	因症状未治疗 (TrNAS)	10
Test and disease	检查证实疾病 (TeRD)	440
	为证实疾病而检查 (TeCD)	90
Test and symptom	检查证实症状 (TeRS)	3340
	因症状而检查 (TeAS)	3010
Disease and symptom	疾病导致症状 (DCS)	1930
	症状表明疾病 (SID)	300

5 Relation Extraction

Relation extraction is the process of identifying how the given clinical entities are related within the clinical note where they exist. And these relationships contain a lot of clinical semantic knowledge. And these knowledge can then be applied in many fields [16, 17]. For this task, we creatively design a CNN-based model and achieve exciting results. First of all, we identify 12 common relation types. Their names and occurrence frequencies are shown in Table 2.

5.1 CNN-based Model Architecture

As shown in Figure 4, in the training process, the outermost layer of the model is initial input. It is the sentence in clinical notes. The last layer refers the output which is a vector and each value of the vector corresponds the possibility of a relation. Besides these, there are 5 more layers in the model including feature layer, embedding layer, convolution layer, pooling layer and fully connected layer.

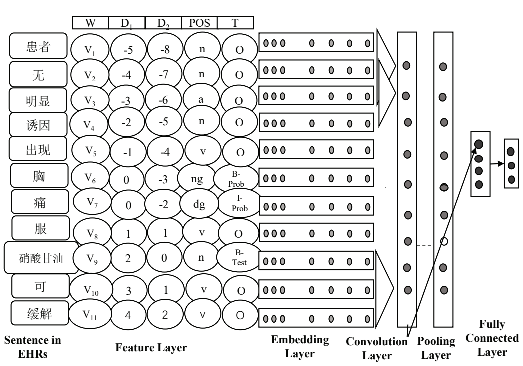

Figure 4: CNN-based Model Architecture

Feature Layer In the feature layer, we introduce five features to represent each word which are word itself (W), distance to clinical entity one (D_1) and entity two (D_2), POS tag (POS) and entity type of the word (T).

- W: the specific words in the sentence
- D_1: the distance between the current word and clinical entity one. Here the clinical entity one and two represent two entities on which we are going to classify relation. And the distance refers to the number of words between the current word and the entity.
- D_2: similar to D_1, D_2 refers to the distance between the current word and clinical entity two.
- POS: POS tag of the current word.
- T: the entity type of the current word. And the type is typically converted into a BIO format.

After obtaining these features, we construct a feature dictionary and all the features are ultimately represented by a numerical matrix.

Embedding Layer In the embedding layer, each feature corresponds to a vector of the embedding feature matrix. Supposing $M^i \in R^{n*N}, i \in \{1, 2...5\}$ is the embedding feature matrix of the ith feature(here n represents the dimension of the feature vector, N represents the number of possible values of the feature or the size of the feature dictionary), then each column in the matrix M^i is the representation of the value of ith feature. Assuming that one hot representation of the jth value of the ith feature is a^i_j, then when the value of the ith feature is j, its vector representation f^i_j is expressed as follows:

$$f^i_j = M^i a^i_j \tag{3}$$

For word embedding, we used word2vec tool [1] to train the word vectors on 55000 clinical notes from a famous medical institute and Q & A data from Chinese medical platform 39 Health [2].

Convolution Layer In convolution layer, we obtain the local features of the sentence by convolution operations. Supposing $x^1 x^2 x^3 ... x^m$ is a feature vector sequence of a sentence with length m, where x^i is the feature vector of the ith word and the length of the filter is c, then the output sequence of the convolution layer is computed as given below:

$$h^i = f(w \cdot x^{i:i+c-1} + b) \tag{4}$$

$f(x)$ is the ReLU function: $f(x) = max(0, x)$. w and b are the parameters we need to train.

Pooling Layer In the pooling layer, we choose the max-pooling to obtain the global feature of each sentence. Not only does this reduce the dimensions of the output, but it still retains the most salient features.

[1] https://code.google.com/archive/p/word2vec/
[2] http://www.39.net/

Table 3: Clinical entities and their frequencies

Entity Type	Number	Entity Type	Number
disease	3110	disease type	220
symptom	24730	test result	3900
test	3230	treatment	4910

Fully connected Layer In the fully connected layer, we use the forward propagation to determine the predicted output. Supposing the output of the pooling layer is a vector $\vec{z}$ whose values come from different filters, then the output o of connecting layer computed as given below:

$$o = W \cdot \vec{z} + b, (W \in R^{[r] \times l}, b \in R^{[r]}) \qquad (5)$$

$[r]$ indicates the number of relationship types.

6 Evaluation

We conducted extensive experiments to evaluate the performance of our model for recognizing clinical entities and extracting relations. In this section, we first introduce our experimental settings. And then report the experimental study results.

6.1 Experimental Settings

All the experiments are done on the real-world clinical notes collected from Beijing Anzhen Hospital. To obtain well-labeled corpus, we first implemented a tool for annotating entities and relations conveniently. And then we invited two medical experts to help annotate the corpus. Specifically, the corpus contains 2200 clinical notes and more than 2039000 words. For clinical entities recognition, we identify 6 clinical entities including " 疾病 "(disease), " 疾病诊断分类 "(disease type), " 自诉症状 "(symptom), " 异常检查结果 "(test result), " 检查 "(test) and " 治疗 "(treatment). Table 3 shows the clinical entities and their occurrence frequencies.

We used the cross validation and chose three metrics: precision (P), recall (R) and F1-score to evaluate all the results. For clinical entity recognition, we combined different features as inputs to evaluate the effect of each one. For relation extraction, we compared the performance of our CNN-based model with two state-of-the-art SVM-based models.

6.2 Experimental Results on Clinical Entity Recognition

For a CRF based model, feature selection is the key to whether the model can achieve good results. To compare the contributions of each feature, we conducted a series of experiments with different features. We started with the model which only use the bag-of-characters feature(W) as our baseline. Table 4 shows the different templates designed with bag-of-characters feature. And Figure 5 shows the performance of different templates.

Table 4: Templates of bag-of-characters feature

Templates	T1	T2	T3	T4
Window size	4	4	5	5
Feature	bigram	trigram	bigram	trigram
Templates	T5	T6	T7	T8
Window size	6	6	7	7
Feature	bigram	trigram	bigram	trigram

It can be observed that for bag-of-characters feature, the templates with bigram features obtained a better performance than templates with trigram features. And the templates with the context window size of 5 achieved the best performance.

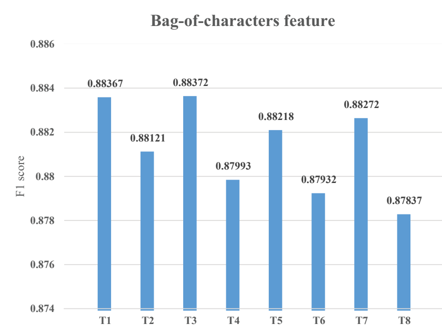

Figure 5: Performance of different templates

Figure 6 describes the performance of the model when POS tag feature(P) and dictionary feature(D) been used. Template 1(T1) is the baseline with only bag-of-characters feature been used. Template 2(T2) to Template 5(T5) respectively added POS tag feature and dictionary feature and used window sizes of 3 to 6. As we can see from the figure 6, with the different size of context window, the templates with the POS tag feature and the templates with the dictionary feature showed the same changing trend and the better performance was generated when the window size is 3. At the same time, the best performance , F1 score of 88.825% was achieved when bag-of-characters feature, POS tag feature and dictionary feature were combined in template 6(T6).

6.3 Experimental Results on Relation extraction

Implementation While implementing our model, we set the word embedding dimension to be 50 and the other 4 feature dimensions to be 5. In other words, the dimension of each word is 70. In convolution layer, we use the combination of filter lengths 3, 4 and 5 together empirically. And we set the number of filters as 100 for every length. Moreover, we use dropout with a probability of 0.50 to prevent overfitting.

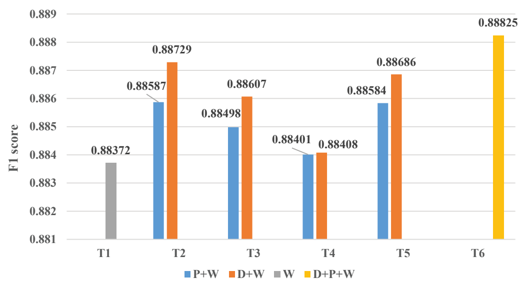

Figure 6: Performance of different feature templates

Table 5: Comparative performance of CNN based model and SVM based models

Model	P	R	F1 score
CNN	87.7%	76.8%	81.4%
Multi-Class SVM	80.3%	62.8%	70.5%
Single SVM	72.0%	75.3%	73.7%

Comparison with featured based models As described before, existing studies for relation extraction problem are mainly based on statistical machine learning methods which heavily depend on manual feature engineering. Here, we compare the performance of our CNN-based model with two state-of-the-art SVM-based models. And we build the SVM classifiers using features defined, respectively, in [8] and [9]. Table 5 shows the comparison of best results obtained by SVM-based models and our CNN-based model.

From the results, we can see that the single SVM model has the lowest precision. But it achieves higher recall than multi-class SVM model since it introduces some new features. And our CNN based model all significantly outperform the two baseline methods, which indicates the effectiveness of our approach.

7 Conclusion

We worked on information extraction on unstructured clinical notes in Chinese EHRs from hospital. Our framework consists of three components: data preprocessing, feature generation and entity and relation extractor. For clinical entity recognition, we propose a novel CRF based model and introduce effective features by leveraging the characteristics of clinical notes and Chinese language. For relation extraction, we utilize CNN to obtain high quality entity-relation facts. A series of experimental results showed that our methods are significantly effective comparing with existing state-of-the-art models.

Acknowledgement

Our work is supported by NSFC(91646202), the National High-tech R&D Program of China (SS2015AA020102), Research/Project 2017YB142 supported by Ministry of Education of The People's Republic of China, the 1000-Talent program, Tsinghua University Initiative Scientific Research Program.

References

[1] G. S. Birkhead, M. Klompas, and N. R. Shah, "Uses of electronic health records for public health surveillance to advance public health," *Annual review of public health*, vol. 36, pp. 345–359, 2015.

[2] P. C. Austin, J. V. Tu, J. E. Ho, D. Levy, and D. S. Lee, "Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes," *Journal of clinical epidemiology*, pp. 398–407, 2013.

[3] M. A. Musen, B. Middleton, and R. A. Greenes, *Clinical Decision-Support Systems*, 2014.

[4] Y. Cheng, F. Wang, P. Zhang, and J. Hu, "Risk prediction with electronic health records: A deep learning approach," in *SDM*. SIAM, 2016, pp. 432–440.

[5] N. Peng and M. Dredze, "Improving named entity recognition for chinese social media with word segmentation representation learning," in *ACL*, 2016, pp. 149–155.

[6] H. He and X. Sun, "A unified model for cross-domain and semi-supervised named entity recognition in chinese social media." in *AAAI*, 2017, pp. 3216–3222.

[7] J. Wang, Z. Wang, D. Zhang, and J. Yan, "Combining knowledge with deep convolutional neural networks for short text classification," in *IJCAI*, 2017, pp. 2915–2921.

[8] A.-L. Minard, A.-L. Ligozat, and B. Grau, "Multi-class svm for relation extraction from clinical reports." in *Ranlp*, vol. 59, 2011, pp. 604–609.

[9] B. Rink, S. Harabagiu, and K. Roberts, "Automatic extraction of relations between medical concepts in clinical texts," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 594–600, 2011.

[10] J.-W. Seol, W. Yi, J. Choi, and K. S. Lee, "Causality patterns and machine learning for the extraction of problem-action relations in discharge summaries," *International journal of medical informatics*, pp. 1–12, 2017.

[11] M. Skeppstedt, "Enhancing medical named entity recognition with features derived from unsupervised methods." in *EACL*, 2014, pp. 21–30.

[12] A. N. Jagannatha and H. Yu, "Bidirectional rnn for medical event detection in electronic health records," in *NAACL*, vol. 2016, 2016, p. 473.

[13] S. K. Sahu, A. Anand, K. Oruganty, and M. Gattu, "Relation extraction from clinical texts using domain invariant convolutional neural network," *arXiv preprint arXiv:1606.09370*, 2016.

[14] L. Yao, Y. Zhang, B. Wei, Z. Li, and X. Huang, "Traditional chinese medicine clinical records classification using knowledge-powered document embedding," in *BIBM*. IEEE, 2016, pp. 1926–1928.

[15] B. He, B. Dong, Y. Guan, J. Yang, Z. Jiang, Q. Yu, J. Cheng, and C. Qu, "Building a comprehensive syntactic and semantic corpus of chinese clinical texts," *Journal of Biomedical Informatics*, vol. 69, pp. 203–217, 2017.

[16] K. Zhao, Y. Zhang, Z. Wang, H. Yin, X. Zhou, J. Wang, and C. Xing, "Modeling patient visit using electronic medical records for cost profile estimation," in *DASFAA*, 2018.

[17] Y. Zhang, X. Li, J. Wang, Y. Zhang, C. Xing, and X. Yuan, "An efficient framework for exact set similarity search using tree structure indexes," in *ICDE*, 2017, pp. 759–770.