Postprint: Structuralization of Chinese Ultrasound Texts and Knowledge Network Construction Methods

Authors: Shang Xiaopu, Xu Wuhuan, Zhao Hongmei, Zhang Runtong, Zhu Shen

Date: 2023-07-26T00:00:00+00:00

Abstract

[Objective/Significance] Ultrasound examination serves as an important basis for assessing patient conditions. Currently, the primary examination data exists in text format. This paper proposes a method for text structuring and knowledge network construction based on ultrasound examination data, aiming to establish a data foundation for further mining of clinical knowledge. [Method/Process] We enhance the application of natural language processing technology in the ultrasound text environment through three main steps: word segmentation, content localization, and structured recognition, to achieve segmentation and labeling of ultrasound texts, and consequently establish a structured knowledge network. [Results/Conclusion] Test results on real data demonstrate that the proposed structuring method for ultrasound examination texts exhibits favorable performance. The method enables automatic construction of structured networks for batch ultrasound texts and can reveal potential knowledge such as hierarchical relationships and attribute structures of structured content within ultrasound texts.

Full Text

Research on Chinese Ultrasound Text Structuring and Knowledge Network Construction Methods

Shang Xiaopu[1], **Xu Wuhuan**[1], **Zhao Hongmei**[1,2], **Zhang Runtong**[1], **Zhu Shen**[1] [1]Department of Information Management, School of Economics and Management, Beijing Jiaotong University, Beijing 100044 [2]Peking University People's Hospital, Beijing 100044

Abstract

[**Purpose/Significance**] Ultrasound examination is a critical basis for patient diagnosis, with most examination data currently existing in text form. This paper proposes a method for text structuring and knowledge network construction based on ultrasound examination data, laying a data foundation for further clinical knowledge mining. [**Method/Process**] We improved natural language processing techniques for the ultrasound text environment through three main steps: segmentation processing, content localization, and structured recognition, enabling segmentation and labeling of ultrasound texts and establishing a structured knowledge network. [**Result/Conclusion**] Real data testing demonstrates that the proposed ultrasound text structuring method achieves favorable performance. This method enables automatic construction of structured networks from batch ultrasound texts and can reflect hierarchical relationships and attribute structures among structured content, revealing latent knowledge within ultrasound texts.

Keywords: Ultrasound text; Natural language processing; Text structuring; Knowledge network

Introduction

Electronic medical records (EMRs) serve as a comprehensive documentation carrier of treatment processes and represent the most important documentation in current medical practice, forming a clinical knowledge repository [1]. Professional medical examination data in EMRs constitute crucial objective data resources for medical big data analytics [2] and provide important data support for evidence-based medicine. As a vital clinical examination method, ultrasound enables rapid and intuitive assessment of conditions in specific body parts. However, unlike most medical imaging examinations and routine laboratory tests, ultrasound results exist solely as text data entered by physicians. As an unstructured data format, text has long posed a significant challenge for precise computer data analysis and knowledge mining. Therefore, analyzing ultrasound examination data requires addressing this fundamental issue. Based on the characteristics of ultrasound data and natural language processing (NLP) technology, this paper proposes a systematic method for text analysis and structuring that can achieve high-precision decomposition and annotation of ultrasound text data with automatic knowledge network construction capability, holding significant scientific importance and application value.

1 Research Status

Natural language processing remains a foundational issue in text mining. Mining hidden knowledge from massive texts can generally be achieved at two levels: first, searching for and extracting specific information for further knowledge mining based on extracted data [3]; second, comprehensively structuring all content into computer-recognizable tokens to build relationship networks and

form knowledge graphs, thereby enabling knowledge mining through reasoning [4]. The former approach works efficiently when mining objectives are clear and the target knowledge is logically understood, but for unknown knowledge or less targeted mining, comprehensive text structuring using the second approach is necessary. Additionally, Chinese and English texts differ in processing: English uses spaces to separate words, while Chinese lacks delimiters between characters. This has led to research focusing on Chinese word segmentation [5] and constructing domain-specific annotated dictionaries [6] that serve as important references for segmenting similar Chinese texts. Current segmentation tools such as Stanford NLP [7], Jieba [8], and Harbin Institute of Technology LTP [9] achieve certain effectiveness for general texts but perform poorly on specialized medical texts, failing to meet technical requirements for further data analysis.

Existing research has addressed medical domain scenarios based on medical text information and knowledge mining methods, including: Chinese-English text analysis and comparison of medical academic literature based on manually constructed corpora [10]; research on the information value of citation contexts in medical academic literature [11]; construction of a Chinese clinical note speculation detection system using bag-of-characters, bag-of-words, character embedding, and word embedding, which demonstrated the importance of segmentation in Chinese clinical NLP [12]; clustering methods for mining medical literature to analyze research directions and hotspots [13]; co-occurrence analysis of clinical variables to facilitate improved high-dimensional propensity score feature selection [14]; a fine-grained semantic description-based medical text retrieval algorithm for internet medical information resources using similarity calculation for content matching [15]; and research on semantic recognition in internet medical documents [16]. However, these studies focused on publicly available medical academic literature or professional materials, with some based on English texts, which differs technically from Chinese clinical record text analysis.

EMR text represents important medical documentation recording examinations, conditions, diagnoses, and other information throughout clinical treatment. Recent domestic research has begun addressing EMR text mining. Studies specifically targeting Chinese EMR texts have achieved segmentation accuracy up to 78.06% using existing segmentation tools [17]. Other research has mined latent semantics in discharge summaries [18], though evaluation was limited to four treatment schemes with limited granularity for clinical application. Additional studies have explored clinical decision support based on EMRs [19-21], focusing primarily on structured/semi-structured data or targeted keyword extraction [22]. Limited research has addressed unstructured medical texts: H. Wang et al. extracted tumor-related information from Chinese liver cancer operation notes using NLP methods [23]; B. He et al. proposed several Chinese clinical text annotation methods from syntactic and semantic perspectives and constructed a comprehensive NLP-based corpus, though with low coverage and annotation efficiency [24].

Transforming EMR free text into a regular form processable by computers is a prerequisite for knowledge mining, often requiring comprehensive application of multiple NLP techniques. Current Chinese medical text structuring commonly converts data into <indicator: indicator value> format based on manually constructed indicator lexicons through information extraction from unstructured text [25-27]. Entity recognition is a crucial objective in text structuring, with studies using deep learning to assign labels to entity attributes [28] and identify disease names [29] and medical event names [30] in EMRs. However, these works did not address inter-entity relationships, which are essential for clinical decision support or data analysis scenarios requiring objective reflection of logical associations between data. Some research has focused on entity relationships in medical records, exploring automatic knowledge graph construction from English EMR texts [31-32]. Chinese EMR knowledge graph construction research has also emerged [33], but remains in early stages with low automation. EMR-based knowledge graphs provide a reliable foundation for clinical knowledge reasoning and diagnosis, with node relationships representing specific semantic relations (e.g., examination-disease, disease-symptom). Ultrasound examination text, as an important EMR component describing ultrasound images of examined body parts, lacks these specific relationships. Therefore, this paper focuses on network structures within ultrasound texts, characterizing "entity-attribute-value" connections and inter-entity hierarchical relationships to constitute an ultrasound knowledge network.

2 Ultrasound Text Data Structuring and Knowledge Network

2.1 Overall Process The proposed ultrasound text structuring and knowledge network construction method consists of three main steps: segmentation processing, content localization, and structured recognition. The segmentation processing stage proposes a word segmentation correction algorithm based on ultrasound text characteristics. Content localization involves text clustering and similarity-based short sentence mapping to achieve semantic content categorization. The structured recognition stage proposes an entity-attribute-value recognition algorithm based on previous processing and maps ultrasound texts to network structures. Inputting batch ultrasound free texts, the method outputs structured data storable in relational databases. Figure 1 [Figure 1: see original paper] illustrates the overall process, with each step detailed in subsequent sections.

2.2 Segmentation Processing As mentioned, widely-used NLP tools supporting Chinese segmentation include Stanford NLP [7-9]. Stanford NLP and Jieba are open-source tools; this study employs Stanford NLP for initial segmentation. While performing well on general language, Stanford NLP's capability for specialized medical texts is suboptimal, though high-quality segmentation is crucial for clinical NLP tasks [12]. A feasible approach involves augmenting professional lexicons and using co-occurrence analysis for automatic lexicon supplementation based on existing NLP tool processing.

For tools like Stanford NLP, out-of-vocabulary words are segmented through specific algorithms. Non-ideal segmentation results exhibit three patterns: (1) "Over-segmentation," where combinable characters/words are split, failing to present fixed collocations; (2) "Under-segmentation," where characters/words that should be separated are treated as a single word; and (3) "Mis-segmentation," where segmentation occurs at inappropriate positions. Ultrasound texts contain numerous abbreviations and special terms, primarily suffering from over-segmentation (Table 1(a)) and mis-segmentation (Table 1(b)). Segmentation quality should be evaluated by whether resulting words correctly express text meaning.

For over-segmentation, this study employs a co-occurrence analysis-based correction method, analyzing adjacent word pairs from initial segmentation results to identify and correct non-ideal segmentation. New words identified through over-segmentation handling also improve mis-segmentation cases. For example, when "strong echo" is over-segmented as "strong" + "echo," correcting "strong echo" as a new word also improves mis-segmentation cases.

Word co-occurrence frequency is calculated as follows: Let S = {W_1, W_2, …, W } represent a data record, where W denotes the i-th word. Word frequency is denoted as Cnt.

Definition 1. The right co-occurrence frequency of word pair (w , w $_1$) is defined as FR(w , w $_1$) = Cnt(w , w $_1$) / Σ Cnt(w , x), where A is the set of all words to the right of w in the text.

Definition 2. The left co-occurrence frequency of word pair (w , w $_1$) is defined as FL(w , w $_1$) = Cnt(w , w $_1$) / Σ Cnt(x, w $_1$), where B is the set of all words to the left of w $_1$ in the text.

Algorithm 1 presents the core pseudocode for the co-occurrence analysis-based segmentation correction algorithm:

```
Input: Adjacent word pairs in text
Output: Candidate new word dictionary Dic
1. for (w , w $_{1}$) do
2.   if Σ  Cnt(w , x) > 1 then
3.     freR $\leftarrow$ Cnt(w , w $_{1}$) / Σ  Cnt(w , x)
4.     if freR $\geq$ C then
5.       dic.append(w , w $_{1}$)
6.   if Σ  Cnt(x, w $_{1}$) > 1 then
7.     freL $\leftarrow$ Cnt(w , w $_{1}$) / Σ  Cnt(x, w $_{1}$)
8.     if freL $\geq$ C then
9.       dic.append(w w $_{1}$)
10. Delete repeated words in Dic
11. Segmentation again with Dic
```

Algorithm 1 input word pairs (w , w $_1$) follow these rules: co-occurrence analysis is skipped if words are separated by punctuation; steps 2 and 6 require

word occurrence > 1 for co-occurrence statistics because word pairs appearing only once yield frequency = 1, which doesn't align with new word discovery objectives and mostly represent noise. For professional terms segmented into three or four words, iterative processing addresses this. For example, "intra- and extra-hepatic bile duct" is initially segmented as "liver" + "intra-extra" + "bile duct." First correction yields "intra-extra hepatic," which added to dictionary Dic enables second segmentation as "intra-extra hepatic" + "bile duct," discovering "intra- and extra-hepatic bile duct" as a new word in the second iteration. Experimental analysis shows ultrasound terms are segmented into no more than four words, with minimal new words discovered in third iteration, thus iteration count is set to 3.

Based on experimental data analysis, threshold C is set to 0.9, meaning combined words "W W $_1$" with right or left co-occurrence frequency ≥ 0.9 are candidate new words. Analysis shows this threshold filters most interference while retaining sufficient new words, yielding domain dictionary Dic for ultrasound texts.

2.3 Content Localization

2.3.1 Text Clustering Text clustering relies on inter-text similarity, requiring similarity matrix calculation for clustering implementation, which enhances subsequent entity, attribute, and value recognition accuracy. Medical texts use relatively professional and direct expressions. While physicians may use different vocabulary to describe the same condition, polysemy is rare, making lexical similarity sufficient for content similarity assessment. This study uses Hamming distance to evaluate similarity between ultrasound reports, clustering them via spectral clustering such that texts within clusters share high linguistic similarity while inter-cluster similarity remains low. Given lengthy records (200-300 characters) and large data volumes, SimHash dimensionality reduction [34-35] is applied before spectral clustering into K clusters.

Algorithm 2 presents the ultrasound text clustering pseudocode:

```
1. foreach Record d do
2.   s  $\leftarrow$ Fingerprint(Record)
3. foreach s  do
4.   d(s$_{1}$, s$_{2}$) $\leftarrow$ HammingDistance(s , s )
5.   sim(s , s ) = 1 - d(s , s ) / hashBits
6.   M.append(sim)
7. SpectralClustering(M, K)
```

Cluster number parameter K was tested at values 3, 4, 5, 6, and 7. Analysis of 4,000 records showed K = 5 effectively separated most dissimilar EMRs while facilitating subsequent experiments, thus this clustering scheme was adopted.

2.3.2 Similar Short Sentence Localization Mapping Following similarity clustering, this step implements relative position mapping of short sentences across ultrasound texts within each category. If records a and b contain similar numbers of short sentences, we establish mapping relationships between a[x] and b[y] to identify descriptions of the same phenomena across different ultrasound texts. Hamming distance is again used to evaluate similarity between ultrasound report short sentences.

Algorithm 3 outlines the short text relative position mapping:

1. Segment short sentences by punctuation boundaries for each record
2. Select the record with the most short sentences as the first record
3. Calculate similarity sim(s_1 , s) between the j-th short sentence of record i and the m-th short sentence of the first record, where i = 2, 3, …, n
4. Extract the short sentence from record i with highest similarity to the m-th short sentence of the first record
5. Process all short sentences in the first record similarly to obtain a similar short sentence mapping table

This algorithm matches the most similar short sentences across ultrasound examination texts, creating a mapping table that identifies and locates short sentences describing identical semantic content across different cases, laying groundwork for entity, attribute, and value recognition.

2.4 Structured Recognition Based on content localization, the proposed algorithm labels segmented content as "entity, attribute, value," establishing a hierarchical ultrasound knowledge network.

2.4.1 Entity, Attribute, and Value Recognition Entities and attributes, as relatively objective description objects, use relatively fixed terminology in ultrasound texts. Values, as specific quantitative or qualitative content for entities and attributes, exhibit richer variation. Due to Chinese writing conventions, "values" typically appear at short sentence ends as numbers or text, though some "character values" precede attributes (e.g., "round/anechoic"). This study identifies entities, attributes, and values based on patterns of fixed versus variable terms within mapped short sentence groups.

Algorithm 4 outlines entity, attribute, and value recognition:

1. Count short sentence frequencies within groups, selecting the most frequent as the standard sentence
2. Select sentences with sim(s, s) > 0.5 to form a set (filtering noise)
3. Segment sentences in the set
4. Compare each sentence's segmentation with subsequent sentences' segmentations; if the latter is a subset of the former, remove the latter sentence
5. Segment remaining set S sentences, counting each word's frequency Cnt(w) and relative frequency f; words with maximum Cnt(w) and f $\geq$ 0.8 are considered entities

6. Locate entity position o in each short sentence; if o+1 is the final position, proceed to (7), otherwise proceed to (8)
7. If Cnt(w $_1$) $\leq$ P, it's a value; otherwise it's an attribute
8. If Cnt(w $_1$) $>$ P or (Cnt(w $_2$) $>$ Q and o+2 $\neq$ e), it's an attribute; otherwise it's a value

Parameters P and Q are set to (number of entity-containing short sentences in S) / 2 for optimal recognition performance based on experimental data.

Using similar short sentence mapping results, entity extraction is performed for each similar short sentence group. The most frequently occurring short sentence serves as the extraction standard. sim(s, s) $>$ 0.5 filters noise. Based on custom dictionary Dic from Section 2.2, each sentence is segmented and compared with subsequent sentences; subset sentences are removed. For example, the set {"liver size morphology normal", "liver morphology normal", "liver morphology size normal", "liver morphology full", "liver morphology abnormal", "liver morphology slightly full", "liver abnormal state"} yields segmentations A{liver, size, morphology, normal}, B{liver, morphology, normal}, C{liver, morphology, size, normal}. B and C being subsets of A, "liver morphology normal" and "liver morphology size normal" are removed. The remaining set {liver size morphology normal, liver morphology full, liver morphology abnormal, liver morphology slightly full, liver abnormal state} is segmented, and "liver" is identified as the entity with frequency 1.0. In the first sentence, w $_1$ is "size" and w $_2$ is "morphology"; since Cnt(w $_2$) $>$ Q and o+2 $\neq$ e, "size" and "morphology" are attributes, while "normal", "full", "abnormal", "slightly full", and "abnormal state" are values. Table 2 shows example entity extraction results.

2.4.2 Network Structure Mapping Post-recognition words form marked entity, attribute, and value libraries. For abdominal ultrasound, five organs are typically examined: liver, gallbladder, pancreas, spleen, and kidney. Using keywords as delimiters, parallel relationships between different organ descriptions are distinguished, with similar mapped short sentences as parallel relationships within the same organ description. Segmented sentences are mapped to recognition result libraries to obtain marked words, organized into target structured forms. The proposed structured storage format is: (primary entity [, secondary entity] [, attribute] [, attribute value]), capturing both intra-sentence "entity-attribute-value" connections and inter-sentence entity hierarchical relationships. Primary entities are the five fixed examination organs; attributes or values may be empty. Using the example from Section 2.4.1, partial structured storage records include "liver-size-normal", "liver-morphology-normal", "liver-morphology-full", etc. This ultrasound text structuring process forms the foundation for corresponding knowledge networks, with each structured storage record representing a path in the network. Visualization tools (e.g., D3.js) can display the resulting ultrasound knowledge network structure.

3 Data Testing and Analysis

3.1 Data Source and Testing Method Data were obtained from abdominal ultrasound examinations at a large tertiary hospital's ultrasound department, totaling 4,818 records. Data were de-identified, removing patient names, IDs, and visit dates, retaining only the "ultrasound findings" field. Random selection yielded 4,600 training records and 218 test records without overlap. Test data were manually segmented and entity-labeled for performance comparison with the proposed method.

3.2 Segmentation Processing Effect Analysis The 218 test records were processed using the domain dictionary from Section 2.2. Figure 2 [Figure 2: see original paper] compares Stanford NLP preprocessing results with co-occurrence analysis-based correction. Figure 3 [Figure 3: see original paper] shows accuracy, recall, and F1 metrics against manual segmentation standards, calculated as:

- Accuracy = (Number of correctly segmented words) / (Total segmented words) × 100%
- Recall = (Number of correctly segmented words) / (Total standard segmentation words) × 100%
- F-value = Accuracy × Recall × 2 / (Accuracy + Recall) × 100%

Results demonstrate that the co-occurrence analysis-based professional dictionary effectively improves segmentation precision, increasing accuracy by 16% compared to existing tools.

3.3 Content Localization Results and Display Following segmentation, content localization was performed on test data. Algorithm 2 clustered experimental texts. Figure 4 [Figure 4: see original paper] displays similarity results for 50 ultrasound texts, where block (i, j) represents similarity between text i and text j, with darker colors indicating higher similarity.

Within each text category, short sentence localization (Algorithm 3) was performed. Figure 5 [Figure 5: see original paper] shows partial content localization mapping results based on clustering.

3.4 Structured Recognition Result Analysis Based on content localization, Algorithm 4's structured recognition capability was tested, labeling results as entity, attribute, or value categories. Random samples of 10, 50, 100, 150, and 218 ultrasound records were tested, with recognition accuracy shown in Figure 6 [Figure 6: see original paper].

Accuracy correlates with sample size, generally increasing with larger samples within a certain range. Physicians follow specific writing conventions, with relatively fixed entities and attributes recurring across patient records while corresponding values vary more extensively. The proposed recognition approach leverages this pattern, making the pattern more evident with larger text vol-

umes. Different examination objects have relatively fixed entities, showing good recognition performance, while attributes and values present greater complexity.

3.5 Ultrasound Knowledge Network Visualization Based on structured recognition of training data, entity hierarchical relationships were established to construct the ultrasound knowledge network shown in Figure 7 [Figure 7: see original paper], which fully preserves ultrasound examination knowledge for structured storage and provides data support for higher-level intelligent diagnostic decision-making applications.

4 Summary and Outlook

This paper proposes an innovative systematic method for ultrasound text structuring and knowledge network construction, integrating multiple algorithms to achieve automatic structuring and network relationship construction from batch medical examination texts, offering new insights for EMR structuring research. The segmentation stage improved accuracy by 16% through co-occurrence analysis-based domain dictionary construction. Content localization grouped descriptions of the same examination objects at both record and short sentence levels based on text similarity, improving structured recognition precision. Real data testing revealed that entity, attribute, and value recognition accuracy generally increases with sample size, demonstrating good performance on large batches.

Limitations include: (1) Inapplicability to small datasets, as the pattern of fixed attributes/values versus variable values within similar short sentence groups describing the same entity is less evident with insufficient samples, causing Algorithm 4 misrecognition; (2) Parameters P and Q in Algorithm 4 may require adjustment and training for different experimental texts. Addressing these limitations represents future work.

Future research can extend the proposed methods to structure more types of medical texts and construct knowledge networks, evolving from single-type medical text knowledge network construction to panoramic medical text structuring and knowledge network construction, establishing a data governance foundation for fully mining hidden medical knowledge.

References

[1] Chen Yongli, Hong Yi. Review on the application of retrieval language in medical information management and retrieval[J]. Library and Information Knowledge, 2015(3): 72-79. [2] Guo Xitong, Zhang Xiaofei, Liu Xiaoxiao, et al. Data-driven electronic health service management research: Challenges and prospects[J]. Management Science, 2017, 30(1): 3-14. [3] Jiménez P, Corchuelo R. On learning web information extraction rules with TANGO[J]. Information Systems, 2016, 62(12): 74-103. [4] Liu Qiao, Li

Yang, Duan Hong, et al. Survey on knowledge graph construction techniques[J]. Journal of Computer Research and Development, 2016, 53(3): 582-600. [5] Zhang Yi, Li Zhijiang. Chinese word segmentation method based on Gaussian word length features[J]. Journal of Chinese Information Processing, 2016, 30(5): 89-93. [6] Guo Shunli, Zhang Xiangxian. Research on sentiment dictionary construction method for Chinese book reviews[J]. New Technology of Library and Information Service, 2016, 32(2): 67-74. [7] Stanford NLP. The Stanford Natural Language Processing Group[EB/OL]. [2018-06-09]. https://nlp.stanford.edu/. [8] Jieba. Jieba Chinese Word Segmentation[EB/OL]. [2018-04-09]. http://www.oss.io/p/fxsjy/jieba. [9] LTP. Language Cloud[EB/OL]. [2018-04-08]. https://www.ltp-cloud.com/. [10] Wang Lanying, Yong Wenming, Wang Lianzhu, et al. Comparative analysis of English abstracts in Chinese and American medical papers[J]. Science-Technology and Publication, 2011(11): 78-82. [11] Liu Yang, Cui Lei. Research on information value of citation context in literature content analysis[J]. Library and Information Service, 2014, 58(6): 101-104. [12] Zhang S, Tian K, Zhang X, et al. Speculation detection for Chinese clinical notes: Impacts of word segmentation and embedding models[J]. Journal of Biomedical Informatics, 2016, 60: 334-341. [13] Yu Yue, Xu Zhijian, Wang Kun, et al. Biomedical informatics text mining research based on biclustering method[J]. Library and Information Service, 2012, 56(18): 133-136. [14] Finlayson SG, LePendu P, Shah NH. Building the graph of medicine from millions of clinical narratives[J]. Scientific Data, 2014, 1: 140032. [15] Guo Shaoyou, Li Yafei, Liang Yuanyuan. Fine-grained semantic description-based medical text retrieval[J]. Information Studies: Theory and Application, 2015, 38(8): 130-134. [16] Wei Wei, Zheng Du. Research on ADR signal extraction model integrating statistical learning and semantic filtering[J]. Library and Information Service, 2017, 62(5): 115-124. [17] Li Guolei, Chen Xianlai, Xia Dong, et al. Research on Chinese EMR text segmentation method[J]. Chinese Journal of Biomedical Engineering, 2016, 35(4): 477-481. [18] Zhang Ye, Zhang Han, Yin Bincan, et al. Disease prediction model construction based on electronic medical records using support vector machine: A case study of severe acute pancreatitis early warning[J]. New Technology of Library and Information Service, 2016, 32(2): 83-89. [19] Lei J, Tang B, Lu X, et al. A comprehensive study of named entity recognition in Chinese clinical text[J]. Journal of the American Medical Informatics Association, 2014, 21(5): 808-814. [20] Liang J, Xian X, He X, et al. A novel approach towards medical entity recognition in Chinese clinical text[J]. Journal of Healthcare Engineering, 2017, 2017. [21] Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: Towards better research applications and clinical care[J]. Nature Reviews Genetics, 2012, 13(6): 395-405. [22] Li Guolei, Chen Xianlai, Xia Dong, et al. EMR text latent semantic analysis for clinical decision-making[J]. New Technology of Library and Information Service, 2016, 32(3): 50-57. [23] Wang H, Zhang W, Zeng Q, et al. Extracting important information from Chinese operation notes with natural language processing methods[J]. Journal of Biomedical Informatics, 2014, 48: 130-136. [24] He B, Dong B, Guan Y,

et al. Building a comprehensive syntactic and semantic corpus of Chinese clinical texts[J]. Journal of Biomedical Informatics, 2017, 69: 203-217. [25] Zhang Yingli, Xia Xiaoling. Structured information extraction method for unstructured pathological texts[J]. Computer Science, 2016, 43(10): 272-276. [26] Chen Dehua, Feng Jieying, Le Jiajin, et al. Structured processing method for Chinese pathological texts[J]. Journal of Medical Informatics, 2016, 37(4): 54-58. [27] Ding Xiangwu, Zhang Xihua. Medical domain text structuring[J]. Computer Engineering and Design, 2017, 38(10): 2873-2878. [28] Dong X, Chowdhury S, Qian L, et al. Transfer bi-directional LSTM RNN for named entity recognition in Chinese electronic medical records[C]//Dalian, Liaoning, China: 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom). Dalian: IEEE, 2017. [29] Wang Pengyuan, Ji Donghong. Disease name extraction based on multi-label CRF[J]. Application Research of Computers, 2017, 34(1): 118-122. [30] Hou Weitao, Ji Donghong. Medical event recognition research based on Bi-LSTM[J]. Application Research of Computers, 2018, 35(7): 1974-1977. [31] Bean DM, Wu H, Iqbal E, et al. Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records[J]. Scientific Reports, 2017, 7(1): 16416. [32] Rotmensch M, Halpern Y, Tlimat A, et al. Learning a health knowledge graph from electronic medical records[J]. Scientific Reports, 2017, 7(1): 5994. [33] Huang Mengxing, Li Menglong, Han Huirui. Research on entity recognition and knowledge graph construction based on electronic medical records[J/OL]. Application Research of Computers: 1-7[2019-03-12]. http://kns.cnki.net/kcms/detail/51.1196.TP.20181129.1122.011.html. [34] Charikar MS. Similarity estimation techniques from rounding algorithms[C]//Montreal, Quebec, Canada: Proceedings of the thirty-fourth annual ACM symposium on Theory of computing. ACM, 2002: 380-388. [35] Rezaeian N, Novikova GM. Detecting near-duplicates in Russian documents through using fingerprint algorithm Simhash[J]. Procedia Computer Science, 2017, 103: 421-425.

Author Contributions

Shang Xiaopu: Conceived research ideas, designed research framework, drafted manuscript; Xu Wuhuan: Collected, cleaned, and analyzed data; Zhao Hongmei: Collected, cleaned, and analyzed data; Zhang Runtong: Revised final manuscript; Zhu Shen: Conducted experiments.

Note: Figure translations are in progress. See original paper for figures.