Post-print: Segmentation and Entity Recognition of Chinese Electronic Medical Records

Authors: Wang Ruojia, ZHAO Changyu, Wang Jimin

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

Abstract

[Purpose/Significance] Health medical big data constitutes a crucial foundational strategic resource in China. This study's investigation and empirical analysis of Chinese electronic medical record segmentation and entity recognition effectively accomplishes the information extraction task for medical data, holding significant importance for the future development of semantic-level applications of medical big data. [Method/Process] This study first constructed a medical lexicon reaching the scale of 100,000 entries by integrating authoritative vocabularies, official standards, health website data, and other medical supplementary dictionaries. Subsequently, segmentation was performed on electronic medical record fields, comparing the segmentation performance of four models: the jieba tool, jieba with imported dictionary, unsupervised learning, and the AC automaton. Finally, utilizing automatic segmentation and manual annotation results as corpus, electronic medical record entity recognition research based on conditional random fields was implemented, and entity recognition performance across different entity categories and text features was compared to select the optimal template. [Results/Conclusion] Segmentation results demonstrate that the AC automaton achieved the best performance, with an F-score reaching 82%. Entity recognition results indicate that the recognition performance for "examination" and "disease" entities was optimal, while the recognition performance for "symptom" was less satisfactory.

Full Text

Preamble

Word Segmentation and Named Entity Recognition in Chinese Electronic Medical Records

Wang Ruojia[1],[2], Zhao Changyu[1], Wang Jimin[1]

[1]Department of Information Management, Peking University, Beijing 100871
[2]Institute of Ocean Research, Peking University, Beijing 100871

Abstract

[**Purpose/Significance**] Healthcare big data represents a crucial foundational strategic resource in China. This study explores and empirically validates Chinese electronic medical record (EMR) segmentation and entity recognition, successfully completing information extraction tasks for medical data and holding significant implications for future semantic-level applications of healthcare big data. [**Method/Process**] This research first constructed a medical lexicon reaching 100,000 entries by integrating authoritative vocabularies, official standards, health website data, and other supplementary medical dictionaries. EMR fields were then segmented, comparing four models: jieba tool, jieba with imported dictionary, unsupervised learning, and AC automaton. Finally, using automatic segmentation and manual annotation results as corpus, we implemented EMR entity recognition based on Conditional Random Fields (CRF), comparing recognition performance across different entity categories and text features to identify the optimal template. [**Result/Conclusion**] Results demonstrate that the AC automaton achieved the best performance with an F-score of 82%. Entity recognition results indicate that "examination" and "disease" entities were recognized most effectively, while "symptom" recognition was less satisfactory.

Keywords: healthcare data mining; electronic medical record; Chinese word segmentation; named entity recognition; AC automaton; conditional random field

Introduction

Electronic medical records are digital information—including text, symbols, charts, graphics, numbers, and images—generated by medical personnel during healthcare activities using information systems. They represent complete and detailed clinical information resources produced and recorded during residents' medical visits [1]. These resources contain substantial latent knowledge. Mining them can provide clinical decision support for medical staff while transforming healthcare delivery models and improving service efficiency and quality. However, research on healthcare big data in China remains in its early stages, with most EMR information stored as unstructured text that cannot be identified or mined through simple segmentation. Advanced segmentation methods and further medical entity recognition are required to address the characteristics of medical texts: numerous specialized terms, many out-of-vocabulary words, and structured descriptive language.

Consequently, this study begins with Chinese EMR segmentation, collecting medical vocabulary by category to form a medical dictionary, and compares dictionary-based methods, unsupervised learning methods, and hybrid approaches for processing Chinese medical records. We then employ

the Conditional Random Field algorithm to identify five medical entities—disease, symptom, examination, medication, and procedure—completing the information extraction task for EMRs.

Literature Review

2.1 Chinese Word Segmentation Methods

Chinese word segmentation is the most fundamental language processing module in natural language processing. Currently, common Chinese segmentation methods primarily include two types: dictionary-based language matching models and statistical/machine learning-based computational models.

2.1.1 Dictionary-Based Language Matching Models

This method involves matching each word in a dictionary against the document being processed. Common matching approaches include maximum matching, reverse maximum matching, minimum matching, and minimum segmentation methods. These share similar basic principles, primarily using a large collection of compiled dictionary entries to match against processed documents to identify possible segmentation patterns.

2.1.2 Statistical and Machine Learning-Based Computational Models

These methods calculate the probability of whether a segmented fragment constitutes a word group. The most common models are Hidden Markov Models (HMM) and unsupervised segmentation models. Liu et al. [2] used the Viterbi algorithm to label globally optimal role sequences, then employed HMM to identify out-of-vocabulary words and calculate their credibility. Li Zhaofu [3] utilized K-shortest path algorithms to form directed graphs of tested text, reducing algorithmic time complexity and providing new approaches for segmentation.

Since 2014, many scholars have applied segmentation technology to the medical domain. Medical texts contain numerous specialized terms such as drugs, diseases, body organs, and surgical methods. Using common segmentation tools significantly reduces segmentation effectiveness and recognition rates in EMRs, necessitating specialized recognition measures for medical terminology. Zhang Libang [4] employed semi-supervised learning—including ordered clustering and Expectation-Maximization (EM) algorithms—for EMR segmentation and named entity recognition of objects and medications. Zhang et al. [5] subsequently used unsupervised methods, converting out-of-vocabulary word recognition into an optimization problem between words through information entropy-based goodness measures, solving segmentation results using dynamic programming algorithms. Li Guolei et al. [6] combined dictionary and statistical segmentation algorithms for discharge record processing, conducting latent semantic segmentation of clinical terms and treatment plans in gastric cancer research.

2.2 Entity Recognition Methods

As this study focuses on clinical EMRs, we summarize entity recognition methods in the healthcare domain, which primarily include: dictionary and rule-based methods, and machine learning-based methods.

2.2.1 Dictionary and Rule-Based Medical Entity Recognition

Rule-based medical recognition relies on terminology dictionaries and domain experts. As early as 1995, experts at Columbia Presbyterian Medical Center designed the MedLEE system [7] to extract, structure, and encode clinical information from patient reports for integration with clinical information systems. Nearly 20 years later, MedLEE still achieves good results, largely due to support from large medical controlled terminology dictionaries. The Unified Medical Language System (UMLS) developed by the U.S. National Library of Medicine since 1986 [9] and the Medical Entities Dictionary (MED) created by CPMC [10] provide foundations for rule-based medical entity recognition. Although this method is relatively primitive, it remains widely used due to convenient open-source tool support such as MetaMap [9], MedEx [10], and cTAKES [11], though these tools only target English texts, as authoritative Chinese medical dictionaries are scarce and rule design cannot cover all special cases.

2.2.2 Machine Learning-Based Medical Entity Recognition

Due to the expert dependency of rule acquisition, researchers increasingly focus on machine learning-based medical entity recognition. Y. Li and S.L. Gorman [12] used 9,679 English clinical reports to build HMM models, identifying the occurrence sequence of different report modules (chief complaint, allergies, family history, past surgical history, etc.). Wang Pengyuan and Ji Donghong [13] analyzed compound disease issues in English medical records, constructing a multi-label CRF model. For Chinese EMRs, Ye Feng et al. [14] used Conditional Random Fields (CRF) with part-of-speech, word formation patterns, word boundaries, and contextual features to identify three named entities—disease, clinical symptom, and surgical operation—in 250 Chinese medical records. J. Lei et al. [15] compared CRF, SVM, Maximum Entropy (ME), and Structured SVM (SSVM) for entity recognition in 400 Chinese admission and discharge records, finding SSSVM achieved the highest F-score (93.51% for admission records, 90.01% for discharge summaries). J. Liang et al. [16] proposed a cascaded Chinese medicine entity recognition method combining SVM and CRF algorithms, achieving 94.2% accuracy for Chinese medicine recognition and 91.7% F-score for Western medicine.

In summary, we find that: (1) In EMR segmentation, despite its growing importance, the unstructured nature and lack of uniform format in medical records make analysis difficult to improve, with strong domain specificity and a general lack of universal analytical methods; (2) In EMR entity recognition, manual segmentation is typically used for case annotation, but different individuals' segmentation standards are difficult to unify, and segmentation granularity sig-

nificantly impacts entity recognition accuracy. This study treats segmentation and entity recognition as two related tasks, first selecting the best segmentation algorithm through comparison, then conducting manual annotation based on automated segmentation results.

Research Design and Methods

3.1 Research Design

The technical roadmap of this study is shown in Figure 1 [Figure 1: see original paper]. The entire research process consists of three steps: data collection, experimental research, and result evaluation. First, we integrated authoritative vocabularies, official standards, health website data, and other supplementary medical dictionaries to build a medical lexicon, while collecting publicly available Chinese EMR data from the internet and preprocessing it. We then conducted experiments on Chinese word segmentation and medical entity recognition. In the segmentation step, we compared four models: jieba tool, jieba with imported dictionary, unsupervised learning, and AC automaton. In the entity recognition step, we implemented CRF-based EMR entity recognition using automatic segmentation and manual annotation results as corpus. Finally, we evaluated results by comparing segmentation outcomes and entity recognition performance across different text features to select the optimal template.

3.2 Data Sources

The data used in this study includes experimental data for segmentation evaluation and entity recognition training/testing, and dictionary data for building segmentation lexicons.

3.2.1 Experimental Data

Due to data security concerns, our EMR data came from clinical physician skill examination simulation questions. The first station of this exam involves history taking and case analysis. Each case consists of ten components: patient gender, age, chief complaint, summary (present illness, past history, personal history), examination (physical examination, auxiliary examination, laboratory tests), diagnosis, diagnostic basis, differential diagnosis, further examination, and treatment principles. Table 1 shows an example case. We collected 100 similar cases from the internet (90 Western medicine cases and 10 Traditional Chinese Medicine cases) for segmentation and entity recognition experiments.

3.2.2 Dictionary Data

Since we required a segmentation dictionary with category labels, the first step in dictionary construction was defining categories. UMLS provides the most authoritative definition of medical entities, with 133 semantic types including anatomical structures, biological functions, and physical objects across six major categories [17]. However, EMRs do not involve so many entity types. The 2010

i2b2/VA challenge referenced UMLS semantic types to categorize EMR named entities into three types: Medical Problem, Test, and Treatment [18]. We further subdivided Medical Problem into disease and symptom, and Treatment into medication and procedure, resulting in five medical entity categories: disease, symptom, examination, medication, and procedure.

The second step involved collecting and crawling authoritative vocabularies, official websites, popular health websites, and other supplementary sources for each category: - **Authoritative vocabulary data**: Chinese Medical Subject Headings based on the U.S. National Library of Medicine's MeSH and the Chinese Traditional Medicine Subject Headings - **Official website data**: Basic drug directories, clinical laboratory test item directories from health commission websites, and drug lists from the national food and drug administration - **Health website data**: Clinical terminology dictionaries, test reference values, disease clinical department classifications, and procedure libraries from medical education websites - **Other supplementary dictionaries**: Baidu Baike medical entries and Sogou medical lexicons

The third step involved preprocessing the collected dictionary: (1) labeling entity categories and source vocabularies for each term; (2) removing classification codes and parenthetical supplementary statements or English names; (3) extracting synonyms within brackets as new terms; (4) filtering terms longer than 20 characters for manual review.

Through these steps, we constructed a medical segmentation dictionary containing 136,253 terms: 54,601 disease terms, 3,316 symptom terms, 3,828 examination terms, 57,390 medication terms, and 17,118 procedure terms.

3.3 Segmentation Methods and Principles

This study compared three segmentation methods: jieba Chinese word segmentation, AC automaton, and unsupervised segmentation.

jieba Chinese Word Segmentation: A common Chinese segmentation tool providing word segmentation and keyword extraction [19]. jieba uses "maximum matching" rules for dictionary-based candidate word selection and result return. If GEN(X) represents the candidate word set generated by dictionary search for tested text X, the dictionary-based model can be expressed as [20]:

$$Y = argmax_{Y' \in GEN(X)} P(Y')$$

AC Automaton (Aho-Corasick): A multi-pattern matching algorithm based on Trie tree structure [21], widely used in information retrieval and string matching. The process involves three steps: (1) building a Trie tree for pattern strings; (2) adding failure paths to the Trie; (3) searching text using the automaton. The search process handles two scenarios: when the current keyword matches, the algorithm proceeds to the next node; when it doesn't match, it follows failure pointers until reaching the root node.

Unsupervised Segmentation: This method uses cohesion degree to determine whether a fragment constitutes a word. Generally, when cohesion exceeds a threshold, the fragment is considered a word. Conversely, when cohesion falls below a threshold, the fragment cannot be a word. For adjacent characters a and b in tested documents, we count co-occurrence frequency F(a,b) and individual frequencies F(a) and F(b). If the following condition holds, the two characters can be separated:

$$\frac{P(a,b)}{P(a)P(b)} < \alpha \quad \text{(where } \alpha \text{ is a given threshold } > 1)$$

After initial segmentation, word frequency can be used to filter candidate word sets.

3.4 Entity Recognition Principles and Methods

In natural language processing, Conditional Random Fields (CRF) is a sequence labeling algorithm combining HMM and Maximum Entropy models. It considers not only word and contextual features but also external features like dictionaries, achieving good entity recognition performance. CRF provides conditional probability distributions of output variables given input variables. For labeling tasks, it is simplified to linear-chain CRF where input and output variables share the same structure:

$$P(y|x) = \frac{exp(\sum_k \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_l \mu_l s_l(y_i, x, i))}{Z(x)}$$

where Z(x) is the normalization factor:

$$Z(x) = \sum_y exp(\sum_k \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_l \mu_l s_l(y_i, x, i))$$

In this study, x represents the EMR text sequence variable, y represents entity annotation variables, and P(y|x) is the conditional probability distribution of output sequence y given x. CRF can consider multiple text features through two feature function types: (1) state feature functions s_l(y_i, x, i) where entity categories depend only on current text; (2) transition feature functions t_k(y_{i-1}, y_i, x, i) that consider external features (part-of-speech, case, context) affecting entity categories.

Key features used in this study include: 1. **Part-of-speech features**: Entity labeling correlates with word POS tags (e.g., named entities are predominantly nouns) and neighboring words' POS tags (e.g., disease entities often appear near verbs like "suffer from" or "diagnose"). 2. **Contextual features**: EMR texts have inherent patterns requiring "context window" length selection. For example, in "因骑车进行中被汽车撞倒右颞部着地半小时到急诊就诊", assuming position i with

word "右颞部" and window length 5, the algorithm extracts words at positions i-2, i-1, i, i+1, i+2. 3. **EMR module features**: In our samples, each EMR contains 10 modules (gender, age, chief complaint, summary, examination, diagnosis, etc.). Different entity categories appear with varying frequencies across modules—diseases predominantly in diagnosis/differential diagnosis modules, while medications and procedures appear in treatment modules. 4. **Position in module**: Relative position within a module may reflect entity category. For example, in chief complaint modules, most sentences follow "symptom + degree" patterns (e.g., "abdominal pain 2 hours, vomiting 3 times"). Calculating position order and total word count yields a normalized value in [0,1] indicating relative position.

We used CRF++ tool to implement CRF-based entity labeling, requiring training files, test files, and template files. Figure 3 [Figure 3: see original paper] shows a training file example with columns for EMR text, POS tags, module identifiers, relative positions, and entity categories using BMES tagging (B=beginning, M=middle, E=end, S=irrelevant). Table 2 shows module name-symbol mappings.

CRF++ template files describe context information and feature selection. We primarily used unigram templates (Table 3), where features are defined as %x[row,col] with row indicating relative line position and col indicating relative column position.

3.5 Evaluation Methods

Evaluation uses recall, precision, and F-score metrics for both segmentation and entity recognition:

Segmentation Metrics: - Segmentation Recall = (Correctly segmented words by algorithm) / (Total words in manual segmentation) × 100% - Segmentation Precision = (Correctly segmented words by algorithm) / (Total words segmented by algorithm) × 100% - Segmentation F-score = 2 × Recall × Precision / (Recall + Precision) × 100%

Entity Recognition Metrics: - Entity Recognition Recall = (Correctly recognized i-class entities) / (Total i-class entities in manual annotation) × 100% - Entity Recognition Precision = (Correctly recognized i-class entities) / (Total i-class entities recognized by algorithm) × 100% - Entity Recognition F-score = 2 × Recall × Precision / (Recall + Precision) × 100%

Higher recall and precision indicate better performance, though they are typically mutually exclusive—improving one often reduces the other [22]. The F-score comprehensively evaluates algorithm performance, with higher values indicating better results. Manual segmentation and annotation served as the gold standard through multi-person cross-validation.

Results

4.1 Segmentation Results

We compared four strategies: jieba, jieba with user dictionary, unsupervised learning, and AC automaton. Segmentation principles are shown in Table 4 . Results (Table 5) show AC automaton achieved the highest F-score (82%), followed by jieba with medical dictionary (74%), unsupervised learning based on information entropy (69%), and plain jieba (66%).

Several factors explain these results: 1. **Limited EMR quantity**: Our sample of 100 EMRs was insufficient for unsupervised learning and non-dictionary algorithms, yielding lower recall and precision. Unsupervised learning requires larger samples for multi-boundary entropy analysis, while dictionary-based algorithms depend on dictionary quality rather than sample size. 2. **Sensitivity to measurement units**: EMRs contain numerous units (e.g., ℃ for temperature, U/L for medication dosage, mol/L for cell counts). Dictionary-based algorithms show higher sensitivity and better unit distinction. Unsupervised learning tends to separate "/" from preceding units (e.g., splitting "U/L" into "U" and "/L"). 3. **Strong medical terminology specialization**: Patient histories and treatment plans involve diverse medical terms with low frequency, reducing unsupervised learning effectiveness. Combining dictionary and statistical methods significantly improves segmentation quality.

4.2 Entity Recognition Results

Using stratified sampling, we separated Chinese and Western medicine cases, allocating 70% for training (14,664 words) and 30% for testing (6,028 words). Overall performance is shown in Table 6 with an F-score of 82.9%.

4.2.1 Impact of Entity Categories

Different categories affect recognition results. By F-score (Figure 4 [Figure 4: see original paper]), "examination" and "disease" performed best, followed by "medication" and "procedure," while "symptom" performed poorly. Comparing precision and recall (Figure 5 [Figure 5: see original paper]), all categories showed higher precision than recall, with "medication" showing the most pronounced difference (94% precision, 63% recall). "Symptom" had the lowest recall (62%) despite 88% precision, resulting in suboptimal F-score. "Examination" showed balanced performance (~85% for both metrics).

4.2.2 Impact of Features

Feature selection significantly affects CRF training. We analyzed single features, feature pairs, and all features to identify optimal template settings.

First, to evaluate position features, we compared module-only versus module+position features. Results (Table 7) showed position features not only failed to improve but substantially reduced precision and recall, leading us to exclude them.

Single-feature results (Table 8) showed contextual and module features achieved higher F-scores, with contextual features reaching 90% precision.

Pairwise feature results (Table 9) showed POS+module combination achieved the highest F-score, even outperforming the all-features model, demonstrating that more features are not necessarily better.

Discussion and Conclusion

5.1 Research Findings

Responding to healthcare big data application needs, this study segmented and recognized entities in EMR data, yielding the following results:

1. **AC automaton achieved the best performance with 82% F-score**. This algorithm integrates dictionary and statistical methods, leveraging extensive medical dictionaries for specialized term sensitivity while using statistical methods to discover out-of-vocabulary words. The limited EMR quantity, measurement unit sensitivity issues, and strong medical terminology specialization hindered unsupervised learning and open-source tools. For highly specialized domains, dictionary integration substantially improves segmentation.
2. **CRF-based EMR entity recognition achieved 82.9% F-score**. For different entity categories, "examination" and "disease" performed best due to strong formatting characteristics—examination items are highly similar across EMRs, and disease entities primarily appear in diagnosis/differential diagnosis modules. "Symptom" recognition was less effective due to colloquial descriptions and appearance across multiple modules. Text position features did not improve model performance, possibly due to our numerical representation causing machine misinterpretation. Feature selection quality, not quantity, is key to accurately expressing text semantics.

5.2 Research Limitations

We encountered several challenges:

1. **Dictionary screening, merging, and construction**: Besides official authoritative data, we used online medical dictionaries that expanded vocabulary but lacked guaranteed accuracy. Authoritative medical vocabularies required tedious preprocessing prone to errors. For example, in ICD-10, most bracketed content represents replaceable synonyms, so we extracted bracketed terms as new words. While effective for terms like "Listeriosis [李司忒氏菌病]", this approach incorrectly extracted "typical" from "Medium [typical] cholera." ICD-9-CM contains many auxiliary coding terms like "excluding" and "code also" that are not actual procedure terms, creating processing difficulties.

2. **Manual annotation rule development**: The lack of Chinese EMR annotation corpora required our team to annotate manually. Medicine is highly specialized—while our team included a medical undergraduate, their experience differed from clinicians. Rule development required consulting physicians on ambiguous terms (e.g., whether "fracture," "shock," or "drug allergy" are diseases or symptoms) and detailed discussions on segmentation granularity (e.g., whether "自觉头痛" should be one or two words, or how to segment "心、肺、腹未见异常").

3. **Medical entity modification issues**: Even correctly recognized entities may have semantic differences. Temporal components like "3 weeks ago" in "3 weeks ago pharyngeal discomfort" indicate symptom duration and severity. Uncertainty markers like "?" in "Pulmonary tuberculosis (infiltrative type? chronic fibrous cavitary type?)" indicate diagnostic uncertainty. Most importantly, negation components like "no" in "no scleral icterus" significantly impact diagnosis—recognizing only "icterus" could misrepresent patient health status and adversely affect subsequent data mining.

5.3 Future Directions

Future research will address these limitations by improving dictionary quality, consulting clinicians for annotation rules, and defining/extracting medical entity modifications. Two additional directions include:

1. **Semi-automated entity annotation**: Manual annotation consumed substantial time and effort. Future work will explore applying medical dictionaries to annotation processes or building multi-classifier collaborative training systems to minimize manual workload.

2. **Entity relationship definition and extraction**: EMR entities are not isolated but interrelated, representing core medical knowledge. Beyond named entity recognition, we must define relationships between different entities to construct medical domain knowledge graphs.

References

[1] National Health Commission. Electronic Medical Record Application Management Specifications (Trial) [EB/OL]. [2018-02-20]. http://www.nhpc.gov.cn/yzygj/s3593q/201702/22bb252

[2] Liu Qun, Zhang Huaping, Yu Hongkui, et al. Chinese lexical analysis using cascaded hidden Markov model [J]. Journal of Computer Research and Development, 2004, 41(8): 1421-1429.

[3] Li Zhaofu. Research and implementation of Chinese word segmentation algorithm based on K-shortest path [D]. Harbin: Harbin Engineering University, 2009.

[4] Zhang Libang. Chinese EMR segmentation and named entity mining based on semi-supervised learning [D]. Harbin: Harbin Institute of Technology, 2014.

[5] Zhang Libang, Guan Yi, Yang Jinfeng. Chinese EMR segmentation based on unsupervised learning [J]. Intelligent Computer and Applications, 2014(2): 68-71.

[6] Li Guolei, Chen Xianlai, Xia Dong, et al. Latent semantic analysis of EMR text for clinical decision making [J]. New Technology of Library and Information Service, 2016, 32(3): 50-57.

[7] Friedman C, Hripcsak G, DuMouchel W, et al. Natural language processing in an operational clinical information system [J]. Natural Language Engineering, 1995, 1(1): 83-108.

[8] Sevenster M, Van O R, Qian Y. Automatically correlating clinical findings and body locations in radiology reports using MedLEE [J]. Journal of Digital Imaging, 2012, 25(2): 240-249.

[9] MetaMap. A Tool For Recognizing UMLS Concepts in Text [EB/OL]. [2018-08-18]. https://mmtx.nlm.nih.gov/.

[10] Xu H, Stetson P, Doan S, et al. MedEx: a medication information extraction system for clinical narratives [J]. Journal of the American Medical Informatics Association, 2010, 17(1): 19-24.

[11] Savova G K, Masanz J J, Ogren P V, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications [J]. Journal of the American Medical Informatics Association, 2010, 17(5): 507-513.

[12] Li Y, Gorman S L. Section classification in clinical notes using supervised hidden Markov model [C]//Proceedings of the 1st ACM International Health Informatics Symposium. Arlington, VA, USA: ACM, 2010: 744-750.

[13] Wang Pengyuan, Ji Donghong. Disease name extraction based on multi-label CRF [J]. Application Research of Computers, 2017, 34(1): 118-122.

[14] Ye Feng, Chen Yingying, Zhou Gengui, et al. Intelligent recognition of named entities in electronic medical records [J]. Chinese Journal of Biomedical Engineering, 2011, 30(2): 256-262.

[15] 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.

[16] Liang J, Xian X, He X, et al. A novel approach toward medical entity recognition in Chinese clinical text [J]. Journal of Healthcare Engineering, 2017(2): 1-16.

[17] UMLS. Current semantic types [EB/OL]. [2018-02-20]. https://www.nlm.nih.gov/research/umls/META3_

[18] Uzuner Ö, South B R, Shen S Y, et al. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text [J]. Journal of the American Medical Informatics Association, 2011, 18(5): 552-556.

[19] Jieba Chinese word segmentation [EB/OL]. [2018-02-20]. https://github.com/fxsjy/jieba.

[20] Shen Xiangxiang, Li Xiaoyong. Improving Chinese word segmentation using unsupervised learning [J]. Small Microcomputer Systems, 2017, 38(4): 744-748.

[21] Kong Donglin, Luo Xiangyang, Deng Qihao, et al. Research on intrusion detection system based on AC automaton matching algorithm [J]. Microelectronics & Computer, 2005, 22(3): 89-92.

[22] Li Yuan. Research on word segmentation and feature selection methods in Chinese text classification [D]. Changchun: Jilin University, 2011.

Author Contributions: Wang Ruojia: Conceptualized research framework, collected cases, completed entity recognition literature review, experiments, and writing; Zhao Changyu: Crawled and organized vocabularies, completed segmentation literature review, experiments, and writing; Wang Jimin: Determined research topic, provided guidance and manuscript revision.

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