Postprint: Named Entity Recognition and Knowledge Graph Construction Based on Electronic Medical Records

Authors: Huang Mengxing, Li Menglong, Han Huirui

Date: 2018-10-11T00:00:00+00:00

Abstract

To address the issues existing in research methods for named entity recognition and entity relation extraction in Chinese electronic medical records, this paper proposes an entity recognition and entity relation extraction method based on the combination of bidirectional long short-term memory (BiLSTM) and conditional random field (CRF). The method first employs word embedding technology to convert text into numerical vectors as input to the BiLSTM neural network, then combines the CRF chain structure for sequence labeling to output the maximum probability sequence, and subsequently knowledge-graphs the recognition results. Experimental results demonstrate that this method achieves significant improvements in accuracy, recall, and F-score when performing entity recognition and entity relation extraction on Chinese electronic medical records. The experimental results satisfy the application requirements of systems in clinical practice and provide guidance for research on constructing clinical decision support systems and personalized medical recommendation services.

Full Text

Preamble

Title: Research on Entity Recognition and Knowledge Graph Construction Based on Electronic Medical Records

Authors: Huang Mengxing[1],[2], Li Menglong[1],[2], Han Huirui[1],[2]
[1]. State Key Laboratory of Marine Resource Utilization in South China Sea
[2]. College of Information Science & Technology, Hainan University, Haikou 570228, China

Abstract: Aiming at the problems existing in current research methods for named entity recognition and entity relationship extraction in Chinese electronic medical records, this paper proposes a novel approach that combines bidirectional long short-term memory (BiLSTM) networks with conditional random fields (CRF). The method first employs word embedding technology to convert text into numerical vectors as input to the BiLSTM neural network, then utilizes the CRF chain structure for sequence labeling to output the maximum probability sequence, and finally visualizes the recognition results as a knowledge graph. Experiments demonstrate that this method significantly improves accuracy, recall rate, and F-value for entity recognition and relationship extraction in Chinese electronic medical records. The experimental results meet clinical system application requirements and provide guidance for developing clinical decision support systems and personalized medical recommendation services.

Keywords: entity recognition; entity relation; BiLSTM; knowledge graph

0 Introduction

Electronic medical records (EMR) refer to digital information generated by healthcare professionals using medical institution information systems during patient care, including text, symbols, charts, graphics, and other electronic data that can be stored, managed, transmitted, and reproduced. EMRs represent an invaluable knowledge resource containing large amounts of accurate and detailed patient medical information. Through knowledge extraction from EMRs, researchers can obtain detailed patient medical information to help build clinical decision support systems and, in the future, provide efficient and convenient personalized medical recommendation services for patients or users. Additionally, extracted knowledge can serve as auxiliary information to help physicians overcome knowledge limitations and reduce individual medical errors.

Knowledge graph construction has become a major research focus across various domains. As a semantic network technology introduced by Google in 2012, knowledge graphs aim to improve internet search efficiency by converting real-world relationships between entities into structured representations. Currently, disease prediction research based on knowledge graphs in China' s medical field is still in its early stages, making the construction of medical knowledge systems based on knowledge graphs significant for the development of smart healthcare. By building medical knowledge graphs, we can enhance the accessibility and comprehensibility of medical knowledge.

1 Related Work

By the end of the 20th century, medical informatization had reached a mature stage internationally, with large-scale corpora, established research methodologies, and the development of the Unified Medical Language System (UMLS). In

medical research, entity recognition (NER) and entity relation extraction (RE) in natural language processing (NLP) have remained hot topics and challenging problems. The primary task of entity recognition is to identify existing conceptual terms within the current knowledge framework from electronic medical records, including diseases, symptoms, medications, tests, and treatments. Entity relation extraction aims to discover and establish relationships between two entities, such as disease-symptom relationships and disease-medication relationships. These two stages lay important groundwork for future construction of personalized healthcare service systems.

Research on entity recognition and relation extraction can be broadly categorized into three approaches: dictionary-based methods, rule-based methods, and machine learning methods. Long Guangyu et al. Error! Reference source not found. employed a combination of conditional random fields (CRF) and dictionary-based methods for disease entity recognition, while Wang Ning et al. Error! Reference source not found. used manually constructed rules to identify company names in the financial domain. However, dictionary and rule-based methods rely too heavily on manually constructed resources, resulting in weak generalization and poor portability. Machine learning approaches for entity recognition can be divided into classification-based methods and sequence labeling methods that treat entity recognition as a joint labeling problem for multiple words in a sentence, selecting the label sequence with maximum joint probability. These methods offer strong scalability and adaptability.

Traditional sequence labeling typically uses the "BIO" scheme, where "B" marks the beginning of an entity, "I" indicates continuation, and "O" represents non-entity tokens. In this work, we add an entity category label "C" to create a "BIO+C"scheme, where "C"specifies the entity type. During corpus construction, unified standards are essential. Qu Chunyan et al. [27] developed detailed "Chinese Electronic Medical Record Named Entity and Entity Relation Annotation Guidelines" based on the linguistic characteristics of Chinese EMRs, providing a solid foundation for NLP research in this domain. Li et al. Error! Reference source not found. compared CRF and support vector machines (SVM) for EMR entity recognition, demonstrating CRF' s superior performance. Lample et al. Error! Reference source not found. proposed the LSTM+CRF model, which outperformed CRF alone, with the key advantage of eliminating the need for feature engineering while achieving excellent results using only word vectors.

For entity relation extraction in the medical domain, Uzuner et al. Error! Reference source not found. pioneered the definition of medical entity relations. In other NLP domains, Socher et al. Error! Reference source not found. proposed using recurrent neural networks (RNN) for entity relation extraction. Based on these advances, we employ the graph database Neo4j, which offers the highest adoption rate, to visualize diseases, symptoms, and their relationships graphically, thereby enhancing the convenience of medical services.

2 Corpus Construction

Through analysis of Chinese EMR text characteristics and building upon references [7,8], we developed corresponding annotation standards with the framework shown in [Figure 2: see original paper]. Our corpus data originated from 2,300 Chinese EMRs provided by Haikou Hospital of Traditional Chinese Medicine, covering 15 departments of varying sizes. Before corpus construction, we performed desensitization processing and randomly selected EMRs from different departments for annotation, resulting in 500 fully annotated Chinese EMRs.

Statistical analysis comparing our EMR corpus with publicly available news corpora reveals that entity density in EMRs is significantly higher than in news text. The annotation framework defines entity types and relation types as shown in the EMR annotation specification framework.

3 Knowledge Extraction Model Design

This paper summarizes widely-used entity recognition and relation extraction methods in medical research and proposes a novel framework. Using Neo4j graph database for knowledge management and visualization, our work provides guidance for smart healthcare development. We divide the task into two phases:

a) **Knowledge extraction phase**: The primary tasks are entity recognition and entity relation extraction. Based on the constructed corpus, we use natural language processing techniques to automatically identify entities in EMRs and apply machine learning methods for analysis and extraction to build relationships between medical entities.

b) **Knowledge storage phase**: This phase focuses on storing knowledge in graph structure and visualizing it through Neo4j. By representing disease entities, symptom entities, and medication entities and their interrelationships as a knowledge graph, we create a comprehensive medical knowledge system.

The overall model framework is illustrated in [Figure 1: see original paper].

3.1 Entity Recognition Model

The entity recognition model consists of three layers: The first layer is the word embedding model, which converts text into numerical word vectors for input to the second layer. The second layer is the BiLSTM model, which automatically extracts text features from word vectors for input to the CRF linear layer. The third layer is the CRF model, which performs sequence labeling on the extracted features and considers the entire sentence to achieve globally optimal labeling.

3.1.1 Word Embedding Model Word embedding is a crucial technique that converts text into machine-readable numerical vectors. There are two main cat-

egories: bag-of-words (BOW) models and distributed representations. BOW's representative method, one-hot encoding, is unsuitable as it fails to preserve semantic information. We adopt word2vec, an open-source NLP tool introduced by Google in 2013 that vectorizes all words to quantitatively measure relationships between them. Word2vec offers two training models: CBOW (continuous bag-of-words) and skip-gram. While CBOW uses context to predict the target word, skip-gram uses the target word to predict its context. We employ the skip-gram model to obtain word vectors that preserve semantic features.

However, word2vec alone cannot fully capture syntactic structure features. To better preserve both semantic and structural characteristics, we add a BiLSTM layer during text preprocessing to predict the next character, further representing semantic and syntactic structures. Combining Word2Vec vectors with BiLSTM hidden layer vectors as input to our text feature extraction model, we name this the Tagging Model, shown in [Figure 4: see original paper].

3.1.2 Text Feature Extraction Model BiLSTM Among entity recognition methods, machine learning approaches are widely used due to their strong scalability and adaptability, though several issues remain. Traditional neural networks have fully connected layers without inter-layer connections and cannot capture dependencies between adjacent labels. While RNNs with hidden layer connections solve this problem, they suffer from gradient vanishing, typically assuming current states only depend on nearby previous states to reduce complexity. This leads to long-term dependency problems where gradients tend to vanish or explode after many propagation steps.

LSTM (long short-term memory), proposed by Hochreiter et al. [28] and later improved by Alex Graves, effectively addresses this issue. To leverage contextual information, we extend standard unidirectional RNN to bidirectional LSTM (BiLSTM), containing two network structures: forward propagation (left-to-right) and backward propagation (right-to-left). These are combined through vector concatenation and mapped to k dimensions (where k is the number of label types in the training set) via a connected linear layer to obtain extracted sentence features, denoted as h. The architecture is shown in [Figure 5: see original paper].

3.1.3 Sequence Labeling Model CRF While LSTM can perform sequence labeling, it suffers from label bias problems. The CRF model obtains globally optimal output sequences, outperforming single LSTM. Using a CRF chain structure shown in [Figure 6: see original paper], we label the text features extracted by BiLSTM using the "BIEOS" scheme, where B=beginning, I=inside, E=end, O=outside, and S=single-character entity. Let $C = (c_1, c_2, \cdots, c)$ and $Y = (y_1, y_2, \cdots, y)$ represent the observation and state sequences (input and output) of the CRF chain structure. The conditional probability distribution $P(Y|C)$ is calculated as:

$$P(Y|C) = \frac{1}{Z(C)} \exp\left(\sum_i \sum_j \lambda_j f_j(y_{i-1}, y_i, C, i) + \sum_i \sum_j \mu_j s_j(y_i, C, i)\right)$$

where the transition function $f_j(y_{i-1}, y_i, C, i)$ represents the transition probability between adjacent labels, the state function $s_j(y_i, C, i)$ represents the probability of label y_i at position i, $Z(C)$ is the normalization term, and λ_j and μ_j are corresponding weights. The most likely label sequence y^* is obtained through:

$$y^* = \arg\max_y P(Y|C)$$

3.2 Entity Relation Extraction Model

The primary goal of EMR named entity relation extraction is to identify predefined relationships between two medical entities, such as relations between diseases, symptoms, examinations, and treatments. Current research predominantly treats relation extraction as either an independent multi-classification problem or as a pipeline combined with entity recognition. However, these approaches ignore the association between the two tasks—errors in entity recognition propagate to relation extraction, amplifying the overall error rate.

Traditional pipeline methods first perform entity recognition, then classify relations between recognized entities. While flexible, this approach requires time-consuming manual annotation by domain experts and introduces prior knowledge through extensive labeling, potentially affecting model recognition capability.

Building upon our entity recognition model, we investigate joint entity relation extraction by proposing a novel tagging method that converts position labels into one-hot encoded auxiliary information. We design a joint model called Entity Relation Model (ERM) using "BIEOS" tagging with predefined relations converted to triples: (entity information, entity relation, entity position in relation), such as B-TeRD-1, E-TeRD-2. We consider only cases where one entity belongs to a single triple. The ERM model replaces the CRF layer with a Softmax layer, as shown in [Figure 7: see original paper].

The ERM model consists of four layers: (1) Word embedding layer converting text to vectors; (2) BiLSTM layer for automatic feature extraction; (3) Text features with appended position labels, where vectorized features are combined with entity position tags using one-hot encoding to form triple format; (4) Softmax layer for multi-classification, converting relation classification to a maximum probability problem. The ERM model is illustrated in [Figure 8: see original paper].

3.3 Neo4j Knowledge Storage Model

We adopt a bottom-up approach for knowledge graph construction. Among database systems, Neo4j offers advantages in performance, design flexibility, and development agility, with Cypher language for data manipulation. Using our entity recognition and relation extraction models, we convert outputs into SPO (Subject, Predicate, Object) triples, as shown in . The entire knowledge graph can be viewed as a collection of such triples.

When importing symptom entities into Neo4j, we establish uniqueness constraints for each symptom node since a specific symptom can be caused by multiple diseases. Rotmensch et al. Error! Reference source not found. proposed a symptom weight factor IMPT calculation method based on Naive Bayes and knowledge graphs:

$$\text{IMPT} = \log P(y = 1|x = 1) - \log P(y = 1|x = 0)$$

where IMPT represents the weight factor of a single symptom for a disease, x_i denotes the symptom entity, y_j denotes the disease entity, and values "1" and "0" indicate presence or absence of disease-symptom pairs. A larger IMPT value indicates stronger connection weight between corresponding disease and symptom entities in the knowledge graph.

4 Experimental Results Analysis

4.1 Entity Recognition Experiment Results Analysis

During model training, we incorporate Bootstrapping to expand the dataset iteratively: (a) Train an initial Tagging Model using existing annotated data; (b) Apply the trained model to unannotated corpus to obtain classification labels and probabilities, adding words with probability above threshold to a reliable set; (c) When the reliable set reaches N=500 instances, merge it with the original annotated dataset, retrain the Tagging Model, clear the reliable set, and repeat step (b).

We use 500 annotated EMRs with cross-validation: 100 documents for training and 400 for testing. Comparing different model parameters, we set word embedding dimension to 256, hidden layers to 4 (two per direction), optimizer to Adam, loss function to cross-entropy, learning rate to 0.001, and dropout to 0.3. Entity recognition experiments target four entity types from [Figure 2: see original paper], evaluating accuracy, recall, and F-value. Results are shown in [Figure 9: see original paper].

Compared with standalone BiLSTM, the BiLSTM+CRF combination improves recognition performance. CRF using word vectors as features performs worse than CRF with manually extracted features. B-Tagging Model, which uses Bootstrapping for dataset expansion during training, shows improved recognition effectiveness with enhanced accuracy and recall, demonstrating better gen-

eralization and applicability. Figure 10: see original paper(b) show B-Tagging Model' s accuracy and recall for four entity types, revealing that examination and treatment entities achieve high accuracy due to their special structural and grammatical characteristics, while diseases and symptoms are more prone to classification errors due to similar positions in EMRs.

4.2 Entity Relation Extraction Experiment Results Analysis

For entity relation extraction, experimental parameters match those in Section 4.1. Results in [Figure 11: see original paper] show that while ERM improves accuracy over traditional CRF, its recall and F-value are lower. ERM-T (with position tags) outperforms both CRF and ERM across all three metrics. Figure 12: see original paper(b) present ERM-T' s accuracy and recall for nine entity relation types from [Figure 2: see original paper], showing best performance for TrID (Treatment Improves Disease) and DCS (Disease Conducts Symptom), while TrAD (Treatment Administered for Disease) and TrAS (Treatment Administered for Symptom) perform worst. This occurs because entity pairs within the same position tags are more easily recognized, while larger intervals significantly reduce performance.

4.3 Knowledge Graph Visualization Results Analysis

We transfer structured triple data to a local Neo4j database using Java to construct the knowledge graph, as shown in [Figure 13: see original paper]. In the Neo4j graph storage system, patients, doctors, or users can input symptoms or diseases for analysis and matching against the constructed medical knowledge graph. Based on relationships between symptom/disease entities and medication entities, the system recommends relevant disease knowledge, medications, prevention methods, and dietary advice. This medical knowledge graph system enables both patient self-diagnosis and assists medical staff in accessing disease information, achieving auxiliary medical functions.

Figure 13: see original paper shows search results for the symptom "chest discomfort" using Cypher match statements, while (b) displays results for the disease "common cold."Blue nodes represent disease entities, red nodes represent symptom entities, and connections represent IMPT weight factors. Visualization clearly reveals relationships between single diseases and multiple symptoms, and between single symptoms and multiple diseases.

5 Conclusion

Due to the unique text characteristics of Chinese EMRs and the lack of large-scale, uniformly annotated corpora, research faces numerous challenges. This paper constructs a novel model for named entity recognition by integrating widely-used dictionary-based, rule-based, and machine learning methods with word embedding technology, BiLSTM, and CRF, demonstrating strong performance. Building upon entity recognition, we address the separation problem in

traditional relation extraction by proposing a joint model that improves recognition effectiveness. We also employ Bootstrapping during training to expand the training corpus and enhance model validity.

Despite promising results, several limitations remain. In the relation extraction model, BiLSTM cannot effectively capture relationships between distant entities of different types. Future work could explore attention mechanisms or distributional weighting based on text content to help the model learn relationships between different entity types. Additionally, multi-classification ideas and syntactic tree structures could improve entity association discovery. The current knowledge graph functionality is limited and requires further expansion, such as extending medical knowledge to the corpus to identify new entity types and construct new entity relationships for more comprehensive medical knowledge graph applications.

References

[1] Long Guangyu, Xu Yun. Combining CRF and dictionary based disease named entity recognition [J]. Information Technology and Network Security, 2017, 36 (21): 51-53.

[2] Wang Ning, Ge Ruifang, Yuan Chunfa, et al. Company name identification in Chinese financial domain [J]. Journal of Chinese Information Processing, 2002, 16 (2): 1-6.

[3] Li Dingcheng, Karin K S, Guergana S. Conditional random fields and support vector machines for disorder named entity recognition in clinical texts [C]// Proc of Workshop on Current Trends in Biomedical Natural Language Processing. Stroudsburg: ACL, 2008: 94-95.

[4] Lample G, Ballesteros M, Subramanian S, et al. Neural Architectures for Named Entity Recognition [J]. Computation and Language, 2016: 260-270.

[5] Uzuner O, Mailoa J, Ryan R, et al. Semantic relations for problem-oriented medical records [J]. Artificial Intelligence in Medicine, 2010, 50 (2): 227-234.

[6] Socher R, Huval B, Manning C D, et al. Semantic compositionality through recursive matrix-vector spaces [C]// Proc of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012: 1201-1211.

[7] Yang Jinfeng, Guan Yi, He Bin, et al. Chinese electronic medical record named entity and entity relation corpus construction [J]. Journal of Software, 2016, 27 (11): 2725-2746.

[8] Zhao Fangfang. Research on part-of-speech tagging technology for Chinese electronic medical records [D]. Harbin Institute of Technology, Harbin, 2014.

[9] Rotmensch M, Halpern Y, Tlimat A, et al. Learning a health knowledge-graph from electronic medical records [J]. Scientific Reports, 2017, 7 (1): 1-11.

[10] Yang Jinfeng, Yu Qiubin, Guan Yi, et al. An overview of research on electronic medical record oriented named entity recognition and entity relation extraction [J]. Acta Automatica Sinica, 2014, 40 (8): 1537-1562.

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

[12] Li Hang. Statistical learning methods [M]. Beijing: Tsinghua University Press, 2012.

[13] Hu Zhangrong, Wang Chaobin. Chinese word segmentation algorithm based dictionary and its performance evaluation [J]. Electronic Technology and Software Engineering, 2015 (15): 102-106.

[14] Jia Lirong, Liu Jing, Yu Tong, et al. Construction of traditional Chinese medicine knowledge graph [J]. Journal of Medical Intelligence, 2015, 36 (8): 51-53.

[15] Liu Qiao, Li Yang, Duan Hong, et al. Knowledge graph construction techniques [J]. Journal of Computer Research and Development, 2016, 53 (3): 582-600.

[16] Li Wei, Zhao Dazhe, Li Bo, et al. Combining CRF and rule based medical named entity recognition [J]. Application Research of Computers, 2015, 32 (4): 1082-1086.

[17] Li Lishuang, Guo Yuankai. Biomedical named entity recognition with CNN-BLSTM-CRF [J]. Journal of Chinese Information Processing, 2018, 32 (1): 116-122.

[18] Li Xiaojing, Lin Hailun, Jia Yantao, et al. Online encyclopedia entities tagging method based on page structure and content [J]. Journal of Frontiers of Computer Science and Technology, 2015, 9 (10): 1238-1246.

[19] Dai Xue, Jiang Zhipeng, Guan Yi. Cross-department chunking based on Chinese electronic medical record [J]. Application Research of Computers, 2017, 34 (7): 2084-2087.

[20] Qin Changjiang, Hou Hanqing. Mapping knowledge domain: A new field of information management and knowledge management [J]. Journal of Academic Libraries, 2009, 27 (1): 30-37.

[21] Zhu Muyijie, Bao Bingkun, Xu Changsheng. Research progress on development and construction of knowledge graph [J]. Journal of Nanjing University of Information Science and Technology, 2017, 9 (6): 575-582.

[22] Hirschman L, Sager N. Automatic information formatting of a medical sublanguage [J]. Sublanguage: Studies of Language In Restricted Semantic Domains, 1982.

[23] Zheng Xiaolin, Wang Weiwei, Hu Zhongkai, et al. A Chinese medical knowledge map construction method based on deep learning: China, G06F17//30 [P]. 2017-05-31.

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