Medical Temporal Phrase Recognition Based on BLSTM Networks: Postprint

Authors: Zhang Shunli, Wang Yingjun, Ji Donghong

Date: 2019-01-28T00:00:00+00:00

Abstract

Identifying temporal phrases from medical texts is one of the key technologies in clinical natural language processing. Traditional rule-based and machine learning methods require designing complex rules and extracting features, and most systems adopt a serial approach that leads to error propagation. We propose a neural network architecture based on Bidirectional Long Short-Term Memory (BLSTM) networks that simultaneously identifies temporal expressions and classifies their types: first, character-level vectors of words learned using Convolutional Neural Networks (CNN) and word vectors trained on large-scale biomedical background corpora are combined as input to the BLSTM; then the BLSTM network is used to learn contextual semantic representations of words; and finally Conditional Random Fields (CRF) are used to optimize the labels of the sequence output by the BLSTM. Experiments based on SemEval-2016 Task 12 show that the neural network learning method without any added features improves the F1 score by 3% compared to the highest official score provided for the task.

Full Text

Preamble

Vol. 37 No. 4
Application Research of Computers
ChinaXiv Partner Journal
Accepted Paper

Medical Temporal Phrase Recognition Based on BLSTM Networks

Zhang Shunli[1], Wang Yingjun[1], Ji Donghong[2]
(1. School of Information Engineering, Henan Institute of Science & Technology, Xinxiang, Henan 453003, China;

2. National Network Security College, Wuhan University, Wuhan 430205, China)

Abstract: Extracting temporal phrases from medical texts is a key technology in clinical natural language processing. Traditional rule-based and machine learning methods require designing complex rules and extracting features, and most systems adopt a pipeline approach that leads to error propagation. This paper proposes a neural network architecture based on Bidirectional Long Short-Term Memory (BLSTM) that simultaneously identifies temporal expressions and their types. First, we combine character-level vectors learned by Convolutional Neural Networks (CNN) with word vectors trained on large-scale biomedical background corpora as input to BLSTM. Then, BLSTM networks learn contextual semantic representations of words. Finally, Conditional Random Fields (CRF) optimize the sequence labels output by BLSTM. Experiments based on SemEval-2016 Task 12 show that the neural network approach without any added features improves the F1 score by 3% over the highest official score in this task.

Keywords: temporal phrases; clinical text; LSTM

Extracting temporal information from clinical notes has recently become a hot research topic in biomedical informatics. This information can help medical staff and researchers better understand disease progression patterns, comprehend dynamic medical phenomena, and serves as the foundation for clinical pathway research and intelligent decision support system development [1~3]. Clinical texts are records of medical activities performed by healthcare professionals for patients. Due to the fast-paced and specialized nature of medical work, temporal information in clinical texts exhibits characteristics of non-uniform formatting, non-standardized expression, and interconnection with medical events. These features pose significant challenges for temporal phrase recognition and hinder subsequent utilization of temporal information, making automatic identification of temporal phrases in clinical texts an important research topic that has attracted increasing attention.

For temporal information extraction from English clinical texts, the Clinical TempEval shared task series has had significant impact, having been evaluated three times at SemEval conferences: SemEval-2015 Task 6 [4], SemEval-2016 Task 12 [5], and SemEval-2017 Task 12 [6]. This task primarily focuses on extracting temporal information from hospital-related clinical texts and can be divided into three subtasks: medical temporal phrase extraction, medical event extraction, and medical temporal-event relation extraction. Medical temporal phrase recognition is the first and most critical step in temporal information extraction tasks. The identified temporal phrases serve as the basis for medical temporal-event relation recognition and constitute the core of the entire task.

0 Related Work

Current temporal phrase recognition methods primarily employ rule-based approaches and machine learning methods. Rule-based methods assume that basic temporal phrases in natural language have clear structures and distinct features, and that comprehensive rules can be designed to cover most temporal information. In the general domain, Strötgen et al. [7] designed the HeidTime system based on regular expression patterns to extract temporal expressions, then used linguistic rules to normalize the expressions, achieving the highest F1 score in the TempEval-2 evaluation. Chang et al. [8] designed the SUTIME system, which supports recognition of four temporal types in English text: time, duration, interval, and set, and has been integrated into the Stanford NLP toolkit. Zhong et al. [9] proposed a method called SynTime for extracting temporal expressions, which outperformed existing state-of-the-art methods on Twitter datasets. Due to the diversity and uncertainty of medical temporal phrases in clinical texts, applying general domain rule-based methods requires enormous effort. For example, MayoTime [10] designed over 300 rules to complete medical temporal phrase extraction in the 2012 i2b2 medical information recognition task.

Most current researchers use linear statistical models in machine learning to complete medical temporal phrase extraction tasks. In SemEval-2016 Task 12, most teams adopted such methods. The UTHealth team [11] used a sequence labeler combining Hidden Markov Models (HMM) and Support Vector Machines (SVM) for temporal phrase recognition, achieving the highest F1 score. The LIMSI team [12] and GUIR team [13] used Conditional Random Field (CRF)-based sequence labeling algorithms to extract temporal phrases. The ULISBOA team [14] used an SVM-based sequence labeler. All these teams employed numerous syntactic, lexical, and domain knowledge features without exception, with the UTHealth team also using results from the rule-based SUTIME system [8] as features. Feature engineering construction requires substantial time and effort, and inevitably introduces some inaccurate features during the process, which also affects final recognition performance.

In recent years, the Bidirectional Long Short-Term Memory (BLSTM) network, a variant of recurrent neural networks, has been widely applied due to its ability to better learn contextual semantic features in natural language. Huang et al. [15] extracted spelling features of words as input to BLSTM networks, then used CRF models to optimize the output sequences from BLSTM, achieving 88.83% F1 score on the CONLL2003 chunking task. Experiments demonstrated that BLSTM models can effectively model contextual semantic representations of sentences, and CRF can optimize output labels to obtain a globally optimal label sequence. Ma et al. [16] used CNN models to learn character-level vector representations of words and combined them with word embeddings as input to BLSTM-CRF models, achieving good results on part-of-speech tagging and general named entity recognition tasks without any handcrafted features. Their experiments not only verified the effectiveness of BLSTM-CRF on sequence la-

beling problems but also demonstrated that CNN models can effectively learn spelling features of words. Li et al. [17] used CNN-BLSTM-CRF models for biomedical named entity recognition on Biocreative GM and JNLPBA corpora, achieving state-of-the-art results. In research using deep neural networks for medical temporal phrase recognition, Fries [18] used a simple bidirectional recurrent neural network (vanilla version) for medical temporal phrase extraction in the SemEval-2016 Task 12 evaluation, while Chikka [19] used convolutional neural network architectures for temporal phrase extraction, with recognition performance worse than rule-based and statistical model approaches.

From the overall framework of medical temporal phrase extraction, most systems adopt a pipeline approach, first identifying temporal phrases and then identifying their categories, which to some extent causes error propagation. Lee et al. [11] proposed a method to simultaneously identify temporal expressions and their types, which can avoid error propagation to some degree. Addressing the diversity, relevance, and uncertainty of temporal expressions in clinical texts, we propose a three-layer bidirectional LSTM neural network system for temporal phrase recognition: a) the first layer combines character-level vectors trained by CNN with word vectors trained on large-scale biomedical background corpora as the word vector input layer; b) the second layer uses BLSTM networks to learn contextual semantic representations of words; c) the third layer uses CRF to optimize label output for sequences output by BLSTM networks. Experimental results on the SemEval-2016 Task 12 corpus demonstrate that this model achieves state-of-the-art performance without using any handcrafted features.

1 Model and Algorithm Description

Temporal phrase extraction from clinical texts can be treated as a sequence labeling task. Following previous successful cases, we simultaneously identify temporal expressions and determine their types, using the BIO tagging scheme. For example, "B-Timex-Date" indicates that a word is the beginning of a Date-type temporal expression, while "I-Timex-Time" indicates that a word is inside a Time-type temporal expression.

[Figure 1: see original paper] shows the BLSTM-based neural network architecture proposed in this paper. It consists of three layers: the first layer is the word vector representation layer, the second layer is the Bidirectional LSTM (BLSTM) layer, and the third layer is the tag output layer. First, sentences are split into word sequences. Input words are converted into corresponding word vector sequences by querying a word embedding table, then concatenated with character-level representation vectors of words trained using Convolutional Neural Networks (CNN). The concatenated word vectors serve as input to the bidirectional LSTM layer, which computes and outputs vector representations with word contextual semantics. Finally, the tag input layer uses Maximum Conditional Random Fields (CRF) to perform overall optimization of output tags.

Fig. 1 Neural network architecture based on BLSTM

1.1 Word Vector Representation Layer

The input to the word vector representation layer is words, and the output is vector representations of these words. In this study, word vector representations consist of two concatenated parts: one part obtained by querying a word embedding table, and the other part being character-level representation vectors of words computed using CNN. The word embedding table can be trained on large-scale unlabeled datasets using word2vector [20], or can use publicly available vector sets provided in the domain. Such public datasets, obtained through unsupervised learning on large-scale corpora, have good generalization effects.

Research by Ma et al. [16] shows that convolutional layers in CNN can effectively describe local features of words, and pooling layers can further extract the most representative parts from local features. Therefore, this paper proposes using CNN to extract character-level features (prefixes, suffixes, and spelling) of words in clinical texts, then combining these with word embeddings to improve model performance. The difference between our CNN module and that of Ma and Hovy lies in our initial character vectors: we map different character vectors for uppercase/lowercase letters and punctuation marks.

The CNN structure used in this paper is shown in Figure 2 [Figure 2: see original paper]. For a word of length n, where represents the i-th character in the word, we define as the character vector of this character. Assuming the convolutional neural network window size C is 1, the final vector representation of character is the concatenation of its own character vector and those of one character before and after it, denoted as . The convolution operation uses a fixed-size kernel to convolve over the character vector matrix of the word to extract local features, and finally uses max pooling to obtain the character-level feature vector representation of the entire word.

1.2 Bidirectional Long Short-Term Memory Layer

Bidirectional Long Short-Term Memory (BLSTM) networks are a special type of Recurrent Neural Network (RNN) model that overcomes the gradient vanishing problem in traditional RNN models caused by overly long sequences [21,22]. BLSTM has been successfully applied to many natural language processing tasks including entity recognition, text classification, and machine translation. BLSTM networks are structurally similar to bidirectional RNNs, with the only difference being the use of LSTM units [23] instead of hidden units in bidirectional RNNs. LSTM units use input gates, memory cells, forget gates, and output gates to control whether contextual information is remembered or forgotten.

The structure can be formally represented as follows: for an input vector sequence , determines which new information is stored in the current memory cell, determines which information in the current cell is discarded, and determines

which information will be output to the current . Each LSTM unit takes as input.

[Figure 2: see original paper] Structure of convolutional neural network

To effectively utilize word context information, this paper adopts the BLSTM structure shown in Figure 1. The BLSTM network computes each input sentence in both left-to-right (forward) and right-to-left (backward) orders. After computation, the t-th word in each sentence obtains two different hidden layer vector representations: and . The output of the bidirectional LSTM layer is obtained by concatenating these two vectors, as shown in Equation (2).

1.3 Tag Output Layer

Conditional Random Fields (CRF), proposed by Lafferty et al. [24] in 2001, is an undirected probabilistic graphical model. CRF obtains a globally optimal label sequence by considering relationships between adjacent labels and has shown excellent performance in many sequence labeling tasks in recent years. This paper uses the CRF algorithm to optimize the output results from the BLSTM layer to obtain globally optimal tag output.

For a sentence: , we define P as its scoring result after BLSTM layer computation. P is an n × k matrix, where k is the number of output label types. We define as the probability that the i-th word in the sentence outputs the j-th label. For a predicted sequence , its score can be defined as:

where A is the transition matrix; represents the transition from label i to label j. Here we use the softmax function to define the probability of generating sequence y:

The training data likelihood function is:

During training, we use the AdaGrad stochastic gradient descent model [25] with a learning rate of 0.03 and regularization parameter of 10 . To alleviate model overfitting, Dropout is added to the input/output parts of the BLSTM layer with a value of 0.5.

1.4 Training Parameters

Parameters were adjusted based on development set results. This paper selects the publicly available 200-dimensional Pubmed [26] vector set as the initial word vector lookup table, trained from large-scale biomedical literature and abstracts. Experimental justification for vector set selection is provided in Section 3.2.1. Other vectors are defined as 200-dimensional. The CNN window size is set to 1, and the output vector length is set to 20.

2 Experiments

2.1 Dataset and Evaluation Methods

This paper conducts experiments using the THYME corpus [27] provided by SemEval-2016 Task 12, which consists of pathology reports and clinical notes of cancer patients annotated by the Mayo Clinic. The dataset is divided into training, development, and test sets. The training set contains 293 clinical document cases, while the development and test sets contain 147 and 152 document cases, respectively, containing 3,833, 3,078, and 1,952 temporal expressions.

The evaluation method consistent with the task is used in experiments. In the tables below, P represents precision, R represents recall, and F1 represents F1-measure.

2.2.1 Word Vector Representation

Word vector representation significantly impacts sequence labeling task results. This paper compares three types of word vectors: a) randomly initialized 100-dimensional, 200-dimensional, and 300-dimensional word vectors; b) 100-dimensional, 200-dimensional, and 300-dimensional word vectors trained using the word2vec tool on the THYME corpus provided by the task; c) 200-dimensional Pubmed word vector sets trained by Pyysalo [26] from large-scale biomedical corpora. Experimental results are shown in Table 1 , with the third type of vectors achieving the best generalization results.

Table 1 Results with different choices of word embedding

Word Vector Representation	F1 Score
Random initialization 100-dim	[value]
Random initialization 200-dim	[value]
Random initialization 300-dim	[value]
THYME corpus trained 100-dim	[value]
THYME corpus trained 200-dim	[value]
THYME corpus trained 300-dim	[value]
Pubmed 200-dim vectors	[value]

2.2.2 Neural Network Architecture Layered Experiments

We conduct layered testing of the neural network structure, comparing experimental results to analyze the role of each module in the model. Results are shown in Table 2 . "-CRF" indicates the model without the CRF layer, where label results are output using softmax. "-pretrain" indicates the model does not use the trained Pubmed vector set as initial vectors but uses random initialization instead. "-CNN" indicates the model without the CNN layer, where vectors read directly from the Pubmed vector set serve as word representations. "CNN-BLSTM-CRF" is the final neural network model selected for this paper.

Table 2 Ablation test performance evaluation

Model Variant	F1 Score
-pretrain	[value]
-CRF	[value]
-CNN	[value]
CNN-BLSTM-CRF (our model)	[value]

2.2.3 Dropout Settings

Dropout prevents overfitting in neural network models. This paper tests different Dropout values on the dataset, with results shown in Table 3 .

Table 3 Results with different dropout values

Dropout Value	F1 Score
0.3	[value]
0.5	[value]
0.7	[value]

2.3 Comparison with Existing Work

Based on the above three groups of experiments, this paper selects 200-dimensional Pubmed vectors and Dropout of 0.5 for experiments and compares the results with other outstanding scholars' work, as shown in Table 4 .

Table 4 Performance comparison with previous research

System	F1 Score
CDE-IIITH (Chikka, 2016)*	[value]
Brundlefly (Fries, 2016)*	[value]
UFPRSheffield (Tissot et al., 2015)	[value]
UTHealth (Lee et al., 2016)	79.5
LIMSI-1 (Grouin and Moriceau, 2016)	[value]
CNN-BLSTM-CRF (our model)	[value]

Lee et al. [11] used a sequence labeler combining Hidden Markov Models (HMM) and Support Vector Machines (SVM) with extensive feature engineering (word form, part-of-speech, stemming, and related dictionaries) for temporal phrase recognition, achieving the best F1 score (79.5) in the SemEval-2016 evaluation. Grouin et al. [12] used a combination of Conditional Random Fields (CRF) and a rule-based system (HeidelTime), achieving first place in accuracy in the SemEval-2016 task evaluation but with very unsatisfactory F1 scores. Tissot et

al. [28] designed a rule-based medical temporal recognition system, achieving the best results in the SemEval-2015 evaluation. Among neural network methods, Fries [18] used word vectors trained on two medical corpora as input to a simple bidirectional recurrent neural network (vanilla version) for medical temporal phrase extraction, with both accuracy and F1 scores being very low. Chikka [19] used convolutional neural network architectures for temporal phrase extraction experiments, with results lower than those obtained using SVM.

Through the above comparative analysis, it can be seen that our neural network model achieves the best performance to date without using any handcrafted features.

2.4 Error Analysis

Analysis of experimental results reveals two main types of errors:

a) Prepositions in temporal phrases are often not recognized. For example, in the gold standard temporal phrase "for the past twenty years," our system identifies "the past twenty years," missing the preposition "for." In the training corpus, some temporal phrases with prepositions are annotated with the preposition included while others are not, which affects final recognition performance.

b) Sample size affects neural network system performance. For example, Time-type phrases constitute a very small proportion of the training data, and their recognition accuracy is correspondingly very low. How to address result bias caused by sample imbalance during neural network training is also a problem this paper needs to solve.

3 Conclusion

This paper proposes a deep neural network architecture for extracting temporal phrases from clinical notes. The model uses convolutional neural networks to effectively represent word morphological features, captures sequence contextual semantic information through BLSTM networks, and optimizes label output results using the CRF algorithm. Without using any handcrafted features or medical domain background knowledge, the model outperforms current state-of-the-art systems. The model can also be applied to solve similar sequence labeling problems, such as medical event extraction. In the future, joint learning through multi-task learning can enable the system to achieve better generalization capabilities. For instance, we could train a joint neural network model that simultaneously identifies temporal and event information to obtain better results.

References:

[1] Hao Tianyong, Pan Xiaoyi, Gu Zhiying, et al. A pattern learning-based method for temporal expression extraction and normalization from multi-lingual

heterogeneous clinical texts [J]. Bmc Medical Informatics & Decision Making, 2018, 18 (Suppl 1): 22.

[2] Moharasar G, Tu B H. A semi-supervised approach for temporal information extraction from clinical text [C]// Proc of IEEE Rivf International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future. 2016: [pages].

[3] Liu Zengjian, Tang Buzhou, Wang Xiaolong, et al. CMedTEX: a rule-based temporal expression extraction and normalization system for chinese clinical notes [C]// Proc of AMIA Annual Symposium Proceedings. 2017: 818.

[4] Bethard S, Derczynski L, Savova G, et al. SemEval-2015 task 6: clinical tempeval [C]// Proc of the 9th International Workshop on Semantic Evaluation. 2015: 806-814.

[5] Bethard S, Savova G, Chen Weite, et al. Semeval-2016 task 12: clinical tempeval [C]// Proc of the 10th International Workshop on Semantic Evaluation. 2016: 1052-1062.

[6] Bethard S, Savova G, Palmer M, et al. SemEval-2017 task 12: clinical tempeval [C]// Proc of the 11th International Workshop on Semantic Evaluation. 2017: 565-572.

[7] Strötgen J, Gertz M. HeidelTime: high quality rule-based extraction and normalization of temporal expressions [C]// Proc of International Workshop on Semantic Evaluation. 2010: 321-324.

[8] Chang A X, Manning C D. SUTIME: a library for recognizing and normalizing time expressions [J]. Lrec, 2012, 9 (1): 3735-3740.

[9] Zhong Xiaoshi, Sun Aixin, Cambria E. Time expressions analysis and recognition using syntactic token types and general heuristic rules [C]// Proc of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver Canada: ACL, 2017: 420-429.

[10] Sohn S, Wagholikar K B, Li Dingcheng, et al. Comprehensive temporal information detection from clinical text: medical events, time, and tlink identification [J]. Journal of the American Medical Informatics Association Jamia, 2013, 20 (5): 836-842.

[11] Lee H J, Xu Hua, Wang Jingqi, et al. UTHealth at SemEval-2016 task 12: an end-to-end system for temporal information extraction from clinical notes [C]// Proc of the 10th International Workshop on Semantic Evaluation. 2016: 1292-1297.

[12] Grouin C, Moriceau V. LIMSI at SemEval-2016 task 12: machine learning and temporal information to identify clinical events and time expressions [C]// Proc of the 10th International Workshop on Semantic Evaluation. 2016: 1225-1230.

[13] Cohan A, Meurer K, Goharian N. GUIR at SemEval-2016 task 12: temporal information processing for clinical narratives [C]// Proc of the 10th International Workshop on Semantic Evaluation. 2016: 1248-1255.

[14] Barros M, Lamurias A, Figueiro G, et al. ULISBOA at SemEval-2016 task 12: extraction of temporal expressions, clinical events and relations using ibent [C]// Proc of the 10th International Workshop on Semantic Evaluation. 2016: [pages].

[15] Huang Zhiheng, Xu Wei, Yu Kai. Bidirectional lstm-crf models for sequence tagging [J]. Computer Science, 2015.

[16] Ma Xuezhe, Hovy E. End-to-end sequence labeling via bidirectional lstm-cnns-crf [C]// Proc of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: Association for Computational Linguistics. 2016: 1064-1067.

[17] 李丽双, 郭元凯. 基于 CNN-BLSTM-CRF 模型的生物医学命名实体识别 [J]. 中文信息学报, 2018, 32 (1): 117-122. (Li Lishuang, Guo Yuankai. Biomedical name entity recognition with cnn-blstm-crf [J]. Journal of Chinese Information Processing, 2018, 32 (1): 117-112.)

[18] Fries J. Brundlefly at SemEval-2016 task 12: Recurrent neural networks vs. joint inference for clinical temporal information extraction [C]// Proc of the 10th International Workshop on Semantic Evaluation. 2016: [pages].

[19] Chikka V R. CDE-IIITH at SemEval-2016 task 12: extraction of temporal information from clinical documents using machine learning techniques [C]// Proc of the 10th International Workshop on Semantic Evaluation. 2016: 1237-1240.

[20] Mikolov T, Sutskever I, Chen Kai, et al. Distributed representations of words and phrases and their compositionality [J]. Advances in Neural Information Processing Systems, 2013, 26: 3111-3119.

[21] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult [J]. IEEE Trans on Neural Networks, 1994, 5 (2): 157-166.

[22] Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions [J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 1998, 6 (2): [pages].

[23] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional lstm and other neural network architectures [J]. Neural Networks, 2005, 18 (5): 602-610.

[24] Lafferty, John D, McCallum, et al. Conditional random fields: probabilistic models for segmenting and labeling sequence data [M]// Departmental Papers (CIS). 2001: 282-289.

[25] Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization [J]. Journal of Machine Learning Research, 2011, 12 (7): 257-269.

[26] Pyysalo S, Ginter F, Moen H, et al. Distributional semantics resources for biomedical text processing [C]// Proc of LBM. 2013: 39-44.

[27] 4Th S W, Bethard S, Finan S, et al. Temporal annotation in the clinical domain [J]. Trans of the Association for Computational Linguistics, 2014, 2 (1): 143-154.

[28] Tissot H, Gorrell G, Roberts A, et al. UFPRSheffield: contrasting rule-based and support vector machine approaches to time expression identification in clinical tempeval [C]// Proc of the 9th International Workshop on Semantic Evaluation. 2015: 835-839.

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