2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI) | 978-1-6654-0898-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICTAI52525.2021.00099

Causal Event Extraction using Iterated Dilated Convolutions with Semantic Convolutional Filters

Jianqi Gao[†], Xiangfeng Luo[*†], Hao Wang[†], Zijian Wang[†]
[†]School of Computer Engineering and Science, Shanghai University, Shanghai, China
{gjqss, luoxf, wang-hao, zijianwang}@shu.edu.cn

***Abstract*—Causal Event Extraction (CEE) is a joint extraction task of events and causality, which can help text understanding, event prediction and so on. Recent research has achieved state-of-the-art performance in various Natural Language Processing (NLP) tasks by combining pre-trained models with neural networks. However, ambiguity of event description and long-distance dependence of event causality result in the low accuracy of extractors. In this paper, we propose a model to incorporate in-domain knowledge by taking frequent expression of event causality into account, and use iterated dilated convolutions to expand the perception field of event causality. External causal knowledge is modeled as frequent n-grams with different length, which is used as convolution filters during kernel initialization, enhancing the ability of model to capture the boundary of event description. To obtain long-distance dependence of event causality, we use iterated dilated convolutions to aggregate context from the entire sentence. Experimental results show that our method significantly outperform the baselines with faster convergence speed.**

***Index Terms*—Causal event extraction, Frequent expression of event causality, Iterated dilated convolutions, Convolution filters**

I. Introduction

Causal Event Extraction (CEE) is a subfield of Information Extraction (IE) that automatically extract cause/effect events from plain text, which is of great value for various intelligent applications due to its significant impact on reasoning and decision making. For example, "*The price of raw material rises sharply, leading to a sharp increase in feed cost*". In this sentence, the cause is "*the price of raw material rises*", the effect is "*a sharp increase in feed cost*". Similarly, in "*As Hudson murdered Andrew, he was sent to prison*", the causal relation between the cause "*murder*" and the effect "*send to prison*" can also be extracted. Due to the ambiguity of event description and long-distance dependence of event causality, CEE is still a challenge in NLP.

Conventional methods for CEE can be roughly divided into rule-based methods [1] [2] and statistical methods [3] [4]. Rule-based methods extract causal events using template matching, requiring extensive manual efforts to construct patterns like lexical patterns, syntactic patterns and semantic patterns. The poor generalization ability of rule based methods result in its low recall of CEE. Compared with rule-based methods, statistical methods partially solves the above problems by constructing rich features. However, it requires sophisticated feature engineering, and manual feature engineering may introduce additional noise.

Xiangfeng Luo is the corresponding author.

In recent years, deep learning have become a mainstream of related tasks in the field of NLP, such as relation extraction, sequence tagging and so on. Among these methods, Convolutional Neural Network (CNN) [5] emphasizes n-gram features have proved to be suitable for relation extraction. Besides, pre-trained language models [6] (e.g., BERT) dominate the state-of-the-art results on a wide range of NLP tasks. BERT support other NLP tasks by providing fine-tuning to achieve better performance on various new dataset. However, pre-trained BERT and CNN still have the following shortcomings: 1) BERT is trained from a large amount of non-domain datasets means that it contains a lot of commonsense knowledge, it may not be sufficient for domain specific tasks. 2) Traditional CNN can only extract the features of several adjacent tokens, which means that its ability to perceive context of sentence sequence is limited.

To overcome the above limitations, we propose a novel neural method for CEE as shown in figure 1. Firstly, we use BERT to generate the representation of each token in the sentence. Secondly, to get the knowledge like frequent n-grams of event causality, we use Naive Bayes to obtain the scores of each n-gram. After that, the centroid vectors of the cause/effect n-gram clusters are applied to the weight initialization of iterated dilated convolutions at the beginning of training. Finally, we link the cause/effect events to reduce the impact of incorrect results on the model by using the query attention mechanism, the final contextual representation is fed into BiLSTM+CRF layer. The main contributions of this work are summarized as follows:

- We propose a joint extraction model of event causality based on the pre-trained model BERT, which can fully consider common expression of event causality and global contextual information.
- This work enhances event causality by integrating the frequent n-gram filters into iterated dilated convolutions, capturing semantic features inside events and longer dependence inter-causality across events.
- Experiments conducted on the real-world datasets show that our method significantly outperforms state-of-the-art baselines.

II. Related Work

In this section, we give a brief review of some related works on CEE in different perspectives, including conventional methods and neural networks.

978-1-6654-0898-1/21/$31.00 ©2021 IEEE
DOI 10.1109/ICTAI52525.2021.00099

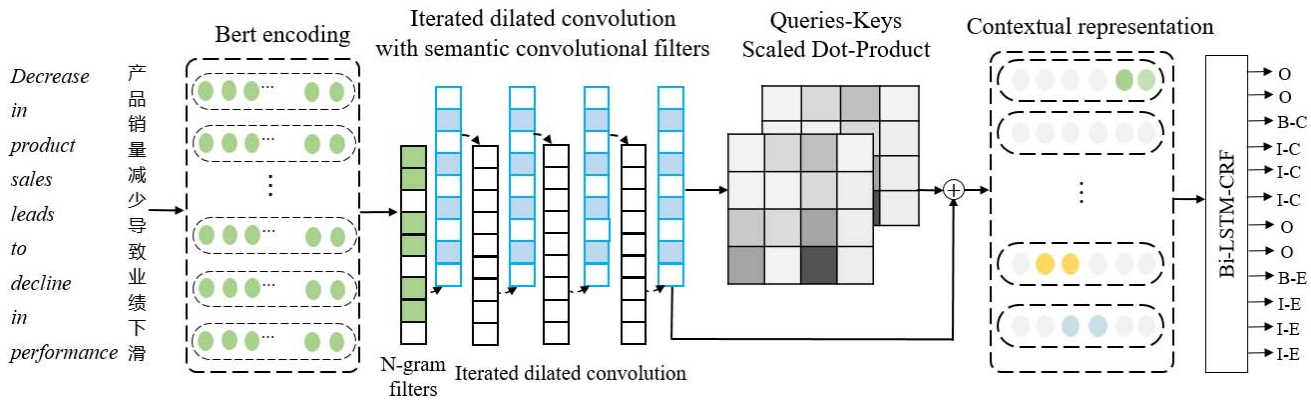

Fig. 1. Overview of the model architecture

A. Conventional Methods

Early studies on causal event extraction mainly focused on rule-based methods, where templates and patterns are handcrafted through manual summary. Garcia and Daniela et al. [1] develop a system, called COATIS that uses the contextual information and heuristic rules to extract event causality. Ittoo et al. [2] present a minimally supervised algorithm that extract both explicit and implicit causal relations from domain specific sparse texts without relying on hand-coded knowledge. However, template matching requires much manual efforts, and the generalization of template matching still needs to be improved. Unlike template matching, statistical learning turns CEE into a classification problem. Blanco et al. [3] improve it by constructing seven types of features. However, this model can only handle explicit causal patterns (e.g.,$< VerbPhrase\ \ Relator\ \ Cause >$). Yang et al. [4] propose a multi-level relation extraction algorithm (MLRE) to recognize all potential causal relations on the basis of dependency/constituency grammar trees. Typically, huge efforts in feature engineering still limit the use of these models.

B. Neural Networks

Recently, deep learning has been widely used in NLP due to its outstanding ability of automatic feature extraction. Among various deep neural networks, CNN has the advantage in capturing semantic features of n-grams and inducing more abstract and discriminative representations of textual inputs [5]. Based on the idea of modeling sentences using CNN, Nguyen et al. [7] prove that convolutional neural networks can significantly improve the performance of relation extraction, which can automatically learn features from sentences with multiple window sizes for filters and pre-trained word embeddings. To reduce the impact of artificial classes, Santos et al. [8] propose the ranking CNN algorithm to minimize novel pairwise ranking loss function. For CEE, Li et al. [9] propose a Knowledge-oriented Convolutional Neural Network (K-CNN) for causal relation extraction. K-CNN consists of a knowledge-oriented channel that incorporates human prior knowledge to capture the linguistic clues of causal relationship, and a data-oriented channel that learns other important features of causal relation from the data. Jin et al. [10] combine the advantage of CNN and LSTM, and propose cascaded multi-structure neural network to extract inter-sentence or implicit causal relations. However, the ability of CNN to perceive contextual information of sentence is limited as convolution kernel only extract local continuous features each time. In response, Yu et al. [11] propose dilated convolutions to capture the contextual information of the text. For dilated convolutions, the effective input width can grow exponentially with the depth. To avoid overfitting, Strubell et al. [12] present Iterated Dilated Convolutional Neural Network (IDCNN), which apply the same small stack of dilated convolutions multiple times, each iterate taking as input the result of the last application. Repeatedly employing the same parameters in a recurrent fashion provide both broad effective input width and desirable generalization capabilities.

III. Our Model

In this section, we will introduce our proposed model, and show how to apply frequent n-grams and iterated dilated convolutions into CEE.

A. BERT Encoder

BERT is a pre-trained transformer network with multi-head attention over 12 (base-model) or 24 layers (large-model) that can be set for various downstream NLP tasks. Here we use it as token encoder. For each token x_t in the input sentence l, BERT will convert it into a fixed-length vector.

B. Iterated Dilated Convolutions with Semantic Convolutional Filters

To improve the model's ability of CEE, and inspired by [9] [13] [14], we can encode semantic features into convolutional filters instead of random initialization. Firstly, we extract cause/effect n-grams from training data, and use Naive Bayes to select important n-grams. Then unsupervised clustering is applied to collect similar cause/effect n-grams into the same clusters, which can better represent semantic causal patterns in the data. After obtaining the n-gram clusters, we

feed the centroid vectors into the center of the filters in the iterated dilated convolutions, the remaining positions are still initialized randomly, which allow the model to learn more useful features itself. The details are given below.

N-gram Selection: Event can be expressed by a phrase with several words, convolution operation can extract semantic features by segmenting the sentence into different chunks. Intuitively, n-gram features (e.g. bi-gram: sales decrease) is a kind of prior knowledge, which can be applied to the stage of convolution initialization. To select effective n-grams, we use Naive Bayes to obtain the scores of each n-gram, the ranking score r of n-gram i is calculated as follows:

$$r = \frac{\left(p_c^i + b\right) / \| \ p_c\|_1}{\left(p_e^i + b\right) / \| \ p_e\|_1} \tag{1}$$

where c is the cause event and e is the effect event, p_c^i is the number of sentences that contain n-gram i in cause c. $\| \ p_c\|_1$ is the number of n-gram in cause c, p_e^i is the number of sentences that contain n-gram i in effect e, $\|p_e\|_1$ is the number of n-gram in effect e, b is a smoothing parameter. At the same time, as the length of event causality in the text is varying in different datasets, we choose the n-grams with the most frequent length in the datasets.

Iterated Dilated Convolutions: In the case of NLP, convolution filters are typically one-dimensional vectors, applied to a sequence of word vectors, or a matrix of filters with same width. To enlarge the perception of sentence sequences, and capture more semantic information. Dilated convolution [11] performs a wider effective input width by skipping over δ inputs at a time, rather than transforming adjacent inputs. Although dilated convolution can capture more contextual information, it increases the depth of the model and introduces more parameters, making the model easy to overfit. To tackle this issue, Strubell et al. [12] proposed a method to reuse the same filter but apply different δ across layers, which can provide both effective input width and desirable generalization capabilities.

The network takes a sequence of vectors x_t as input, suppose that the jth dilated convolution layer of dilation width δ is denoted as $D_\delta^{(j)}$, and $\delta = \{1, 1, 2\}$. The first layer $D_1^{(0)}$ with dilation=1 in the network transform the input x_t to a representation i_t. Then the dilated convolution layer L_c of increasing dilation width are applied to i_t. Beginning with $c_t^{(0)} = i_t$, the stack of layers with recurrence can be represented as follows:

$$c_t^j = r(D_{2^{L_c-1}}^{(j-1)} c_t^{(j-1)}) \tag{2}$$

where $r()$ is the ReLU activation function. Finally, a layer with dilation=1 is added to this stack:

$$c_t^{(L_c+1)} = r(D_1^{L_c} c_t^{(L_c)}) \tag{3}$$

The stack of dilated convolutions is defined as *block* $B(\cdot)$. To incorporate broader context without over-fitting and not introduce additional parameters, iterated dilated convolutions iterative apply B L_b times. Starting with $b_t^{(1)} = B(i_t)$:

$$b_t^k = B(b_t^{(k-1)}) \tag{4}$$

Filter Initialization: Through Formula 1, we can select important cause/effect n-grams. Then each n-gram is embeded into a vector representation using BERT. Since filters in CNN is insufficient to make use of all selected n-grams, we use k-means to bring similar n-grams into a cluster, and the centroid vectors clustered by k-means is applied to filter initialization. Considering that non-causal n-grams exist in the sentence, we do not fully fill the convolution kernel with centroid vectors, and the rest weights of the filters is randomly initialised. The final generated filters is copied to the first layer $D_1^{(0)}$ of iterated dilated convolutions.

C. Query-Key Attention

After obtaining the feature map of the sentence using iterated dilated convolutions with semantic convolutional filters, the query-key attention mechanism [15] is employed to mine internal causal relationship between features. The input consists of query matrix $Q \in \mathbb{R}^{t\times d}$, keys K$\in \mathbb{R}^{t\times d}$ and values V$\in \mathbb{R}^{t\times d}$. In addition, the outputs of iterated dilated convolutions and self-attention structure are concatenated as higher-level contextual representation and fed into BiLSTM+CRF layer.

D. BiLSTM+CRF Layer

LSTM is a variant of recurrent neural network, which is used to solve the problem of gradient vanishing [16]. The LSTM used in BiLSTM [17] mainly consists of three parts, including an input gate i_t, an output gate o_t and a cell activation vectors v_t. BiLSTM uses two LSTM layers to learn the valid characteristics of each token in the sequence based on the past and future context information of the token. Given input contextual representation, we use BiLSTM to model the temporal dependencies of tokens. Conditional Random Field (CRF) [18] can obtain the label of a given sequence in the global optimal chain, and take the interaction between adjacent labels into consideration. Given sentence l and its prediction sequence $y = y_1, y_2, ..., y_n$, CRF score can be obtained by using the following formula:

$$score(l, y) = \sum_{i=1}^{n+1} A_{y_{i-1}, y_i} + \sum_{i=1}^{n} P_{i, y_i} \tag{5}$$

where P_{i,y_i} is the prediction probability of the i-th token in the sentence with label y_i, and the transition probability from label y_{i-1} to y_i is A_{y_{i-1},y_i}.

E. Object Function

The final convergence condition is to minimize the loss function which can be expressed by the following formula:

$$E = \log \sum_{y\in Y} \exp^{s(y)} - \text{score(s, y)} \tag{6}$$

where Y is the set of all possible label sequences in a sentence.

IV. Experiment and Evaluation

In this section, we describe the datasets across different domains and the baseline methods applied for comparison.

A. Datasets

We conduct our experiments on three benchmark datasets, including two Chinese datasets and one English dataset. The first dataset is Chinese Emergency Corpus (CEC), which is an event ontology corpus publicly available[1]. It contains six categories: outbreaks, earthquakes, fires, traffic accidents, terrorist attacks, and food poisonings. The second is Chinese financial dataset, we crawled a large scale of Chinese financial news reports from the internet, such as Jinrongjie[2] and Hexun[3]. This dataset contains a large number of financial articles involving causal relations. The third is English datasets, Since there are few publicly available English dataset for CEE, we get the data from the SemEval-2010 task 8 dataset [19]. The dataset contains 10,717 annotated samples, each sample is a sentence annotated with a pair of entities (e_1 and e_2) and their relationship class. As we only interested in cause/effect relation, we annotate all other relations as other class. We invite three annotators to annotate the data. Each annotator needs to determine whether it is a causal sentence. If it does, they need to annotate which part are the cause or effect. Finally, we annotate 1026 causality instances from CEC corpus, 2270 causality instances from financial corpus, and 1003 causality instances from SemEval-2010 task 8.

B. Experimental Setting

We use the pre-trained model BERT to encode sentences, and set the hyper-parameters of our model as follows: 1) the number of training epochs to 100; 2) the maximum length of sentences to 256; 3) the size of batch to 8; 4) the learning rate for adam to 1×10^{-5}. To prevent over-fitting, the droupout rate of training process is 0.4.

C. Baseline Methods

To prove the effective of our method, we compare our method with the following baselines.

- **BiLSTM+CRF**: This is a basic model for sequence tagging, which uses BiLSTM to mine past and future input features and capture sentence level tag information with CRF.
- **CNN+BiLSTM+CRF** [20]: The model is used for part-of-speech(POS) tagging. They first use CNN to encode each token into character-level representation, and feed them into BiLSTM+CRF layer.
- **CSNN** [10]: The author uses CNN and self-attention to capture features relationship, and the higher-level phrase representations are fed into BiLSTM and CRF layer.
- **BERT+CSNN**: This method is a very effective baseline language model for CEE, which uses pre-trained BERT [6] trained from large-scale unlabeled corpus as input.
- **BERT+IDCSCF**: Our model extract causal event using iterated dilated convolutions with semantic convolutional filters, incorporating with the BERT encoder.

[1]https://github.com/shijiebei2009/CEC-Corpus
[2]http://www.jrj.com.cn/
[3]http://www.hexun.com/

D. Comparison with Baseline Methods

To demonstrate the overall performance of our method, we compare it with other state-of-the-art baselines. Among them, BERT+CSNN is a strong baseline that constructed for comparison. The overall performance in terms of F1 scores is shown in table I.

We can see that BERT+IDCSCF is 1.54%, 2.91% and 2.52% higher than BERT+CSNN on financial, CEC and SemEval2010 respectively, proving the effectiveness of iterated dilated convolutions with semantic convolutional filters produced by pre-trained model.

TABLE I
AVERAGE F1-SCORES OF OUR METHODS ON THREE DATASETS COMPARED WITH OTHER BASELINES.

Model	Financial	CEC	SemEval2010
BiLSTM+CRF	74.75	68.74	73.20
CNN+BiLSTM+CRF	74.31	71.68	74.20
CSNN	74.59	70.61	73.71
BERT+CSNN	76.23	74.61	75.69
BERT+IDCSCF	**77.77**	**77.52**	**78.21**

Compared with CNN-based models, the performance of traditional sequence labeling algorithm BiLSTM+CRF is relatively poor. Although LSTM-based methods can capture both past and future contextual information, and are very effective in sequence labeling such as named entity recognition. However, phrase-level event contain several consecutive words and their expressions are more ambiguous, resulting in the poor performance of BiLSTM+CRF. CNN+BiLSTM+CRF and CSNN achieve better performance than BiLSTM+CRF, these two methods use CNN to extract phrase-level features of several adjacent words, which is crucial for phrase-level event extraction. We can see our model outperforms BERT+CSNN, BERT+IDCSCF integrates prior knowledge like frequent n-grams into the weight initialization of iterated dilated convolutions, which not only enhance the model's ability to distinguish the boundary of intra-event mentions, but also perceive rich contextual information of inter-event causality, and has been experimentally proven to be effective for CEE.

E. Ablation Experiments

We remove attention, SCF (semantic convolutional filters) and BERT in order to see whether each part of the model plays a positive role in CEE. The results of experiments launched on three datasets are shown in figure 2.

The result show that each part of our model is useful for CEE, and the contribution of BERT is most significant for the reason that BERT is a deeper neural network that contains rich semantic knowledge. In addition, the attention mechanism helps the model to filter out some noise, SCF

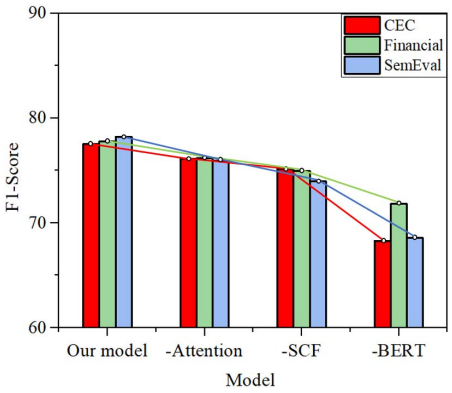

Fig. 2. Ablation Experiments of causal event extraction on three datasets

improve the accuracy and convergence speed of the model for CEE by introducing frequent n-grams of event causality. Iterated dilated convolutions enhance the ability of the model to capture more contextual information. In conclusion, all these layers contribute to the performance of the model.

V. Conclusion

In this paper, we present a novel approach iterated dilated convolutions with semantic convolutional filters for causal event extraction. Our method can effectively integrate frequent n-grams of event causality into iterated dilated convolutions, and improve the accuracy of causal event extraction. We use Naive Bayes to select important n-grams of event causality, and generate the centroid vectors of the cause/effect n-gram clusters. In addition, we apply the centroid vectors to the weight initialization of iterated dilated convolutions at the beginning of training, which can capture semantic features inside events and longer dependence inter-causality across events. The performance of our approach has been experimentally verified on three datasets.

In future work, we will try to build a pipeline model to further improve the accuracy of CEE based on existing research. CEE can be seen as a pipeline task. First, it needs to extract the complete event mentions in the sentence, and then determine which is the cause or effect event. Furthermore, we will explore potential applications of our model in other domain-specific tasks.

Acknowledgment

The research reported in this paper was supported in part by the National Natural Science Foundation of China under the grant 91746203 and the Outstanding Academic Leader Project of Shanghai under the grant No.20XD1401700 and the Ministry of Industry and Information Technology project of the Intelligent Ship Situation Awareness System under the grant No.MC-201920-X01.

References

[1] Garcia D.: COATIS, an NLP system to locate expressions of actions connected by causality links[C]//International Conference on Knowledge Engineering and Knowledge Management, Springer, Berlin, Heidelberg, 1997, pp. 347-352.

[2] Ittoo A, Bouma G.: Extracting explicit and implicit causal relations from sparse, domain-specific texts[C]//International Conference on Application of Natural Language to Information Systems. Springer, Berlin, Heidelberg, 2011, pp. 52-63

[3] Blanco E, Castell N, Moldovan D I.: Causal Relation Extraction[C]//Lrec, 2008.

[4] Yang X, Mao K.: Multi level causal relation identification using extended features[J]. Expert Systems with Applications, vol. 41, no. 16, pp. 7171-7181, 2014.

[5] Kalchbrenner, Nal, Edward Grefenstette, and Phil Blunsom.: A convolutional neural network for modelling sentences. arXiv preprint arXiv:1404.2188, 2014.

[6] Devlin J, Chang M W, Lee K, et al.: Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[7] Nguyen T H, Grishman R.: Relation extraction: Perspective from convolutional neural networks[C]//Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, 2015, pp. 39-48.

[8] Santos C N, Xiang B, Zhou B.: Classifying relations by ranking with convolutional neural networks[J]. arXiv preprint arXiv:1504.06580, 2015.

[9] Li P, Mao K.: Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts[J]. Expert Systems with Applications, vol. 115, pp. 512-523, 2019.

[10] Jin X, Wang X, Luo X, et al.: Inter-sentence and Implicit Causality Extraction from Chinese Corpus[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, Cham, 2020, pp. 739-751.

[11] Yu F, Koltun V.: Multi-scale context aggregation by dilated convolutions[J]. arXiv preprint arXiv:1511.07122, 2015.

[12] Strubell E, Verga P, Belanger D, et al.: Fast and accurate entity recognition with iterated dilated convolutions[J]. arXiv preprint arXiv:1702.02098, 2017.

[13] Li S, Zhao Z, Liu T, et al.: Initializing convolutional filters with semantic features for text classification[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1884-1889.

[14] Wang Z, Wang H, Luo X, et al.: Back to Prior Knowledge: Joint Event Causality Extraction via Convolutional Semantic Infusion[J]. arXiv preprint arXiv:2102.09923, 2021.

[15] Vaswani A, Shazeer N, Parmar N, et al.: Attention is all you need[J]. arXiv preprint arXiv:1706.03762, 2017.

[16] Hochreiter S, Schmidhuber J.: Long short-term memory[J]. Neural computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[17] Lample G, Ballesteros M, Subramanian S, et al.: Neural architectures for named entity recognition[J]. arXiv preprint arXiv:1603.01360, 2016.

[18] Lafferty J, McCallum A, Pereira F C N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data[J], 2001.

[19] Hendrickx I, Kim S N, Kozareva Z, et al.: Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals[J]. arXiv preprint arXiv:1911.10422, 2019.

[20] Ma X, Hovy E.: End-to-end sequence labeling via bi-directional lstm-cnns-crf[J]. arXiv preprint arXiv:1603.01354, 2016.