Intelligent BERT-BiLSTM-CRF Based Legal Case Entity Recognition Method

Mingdong Sun
BeiJing THUNISOFT Information Technology Corporation Limited, Beijing, Chin
sunmd@thunisoft.com

Zhixin Guo
Cyber Security School of Beijing University of Posts and Telecommunications, Beijing China,18874848129@163.com

Xiaolong Deng*
Beijing University of Posts and Telecommunications, Beijing, China
shannondeng@bupt.edu.cn

ABSTRACT

In the past decade, the main natural language processing technologies in the field of artificial intelligence are Word2Vec and ELMO traditional models in the application of intelligent legal systems. For the reason that they are basically one-way training algorithms from left to right and only one-way information is learned, so these traditional models have some disadvantages such as low efficiency and accuracy. In order to identify specific elements in the legal case intelligently, such as time, location, perpetrator, and recipient, and improve the efficiency of case processing, a new entity recognition method using the BERT (Bidirectional Encoder Representations from Transformers) model as the input layer is proposed. The BERT model is a new type of word vector model that relies on context by joint adjusting the bidirectional Transformer in all layers. Basing on BERT model, we proposed a new method comprise BERT, BiLSTM and CRF (Conditional Random Fields) to carry on the intelligent identification of legal case entities. And with abundant experiment result, the better accuracy and efficiency of our method has been proved comparing to traditional models such as Word2Vec.

CCS CONCEPTS

• **Computing methodologies**; • **Artificial intelligence**;

KEYWORDS

Natural Language Processing, Intelligent Legal Affairs, BERT, Entity Recognition

ACM Reference Format:
Mingdong Sun, Zhixin Guo, and Xiaolong Deng. 2021. Intelligent BERT-BiLSTM-CRF Based Legal Case Entity Recognition Method. In *ACM Turing Award Celebration Conference - China (ACM TURC 2021) (ACM TURC), July 30–August 01, 2021, Hefei, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3472634.3474069

*XiaoLong Deng is the Corresponding authors (e-mail: shannondeng@bupt.edu.cn).

ACM TURC, July 30–August 01, 2021, Hefei, China

ACM ISBN 978-1-4503-8567-1/21/07…$15.00
https://doi.org/10.1145/3472634.3474069

1 INTRODUCTION

With the accelerated construction of China's legal society and the rapid development of artificial intelligence technology, the application of artificial intelligence technology in the judicial field is becoming more and more mature. Intelligent labeling and extraction of relevant entities of writing materials such as case records can greatly improve the efficiency of legal case handling, and relevant research is of great significance.

However, the traditional Chinese named entity recognition pre-training model has a vital problem of insufficient extraction ability. And in this paper, the BERT [1] (Bidirectional Encoder) model which is put forward by Google in 2018 will be used by us as the feature expression layer. The pre-training model BERT is a two-way coding representation model taking Transformers as the main framework, which has a very strong textual feature representation capability. Due to the stronger feature extraction ability of BERT and the ability of BiLSTM (Bidirectional Long Short-Term Memory) to obtain long-term sentence memory information and fully consider the context to obtain global semantic information while BiLSTM having a two-way Long Short-Term Memory, We combined BERT with BiLSTM which is a good naming entity recognition model in this paper. Finally, with the state-transition matrix in CRF (Conditional Random Fields) technology to output the globally optimal sequence and we have got some better experimental results. And the main contributions of this paper are as follows.

1) We have used BERT as the input layer to obtain better word vectors. For the reason that BERT model can make full use of the word relationship between contexts, it can effectively extract text features. Aiming at the characteristics of many more referential parts in legal texts, BERT is innovatively used for comprehensive extraction of context features to improve the accuracy of text preprocessing.

2) This article combined BERT+BiLSTM+CRF with entity extraction in actual legal big data, and propose a legal intelligent entity recognition model based on BERT+BiLSTM+CRF. Then it use the People's Daily and the CAIL Law Research Cup related legal text classic data sets and manually sorted and annotated legal transcript text data as the verification data set. The results show that the method proposed in this article can effectively improve the effect of related named entity recognition in legal cases, which is significantly better than Classic models such as Word2Vec and ELMO.

The rest of in this paper are arranged as followed:

The second part introduced related research of the legal text entity recognition and BERT Model. The third part introduced the two pre-training mission model of BERT which are Masked LM

(Masked Language Model) and NSP (Next Sentence Prediction). The forth part introduced the details of the intelligent BERT-BiLSTM-CRF model based legal case entity recognition method adopted by us and the Part Five described the experimental dataset, adopted experiment indicators and analysis to experiment results. And the last part gives the summary conclusion of this paper.

2 RELATED WORK OF BERT AND ENTITY RECOGNITION

The word 'Named Entity Recognition' (NER), which is widely used in natural language processing, was firstly proposed by Grishman R and Sondheim B [2] at the Sixth Message Understanding Conference (MUC-6) in 1996. And NER has always been a hot topic in NLP research field. The earliest rule-based and dictionary-based methods have a high cost of recognition and induction not only, but also have low recognition efficiency. In the traditional machine learning method stage, The NER techniques mainly include Hidden Markov Models (HMM) proposed by Bikel D M, Schwartz R and Weischedel R [3] in 1999, Decision Trees proposed by SekineS [4] in 1998, Maximum Entropy Models (ME) proposed by A. Borthwick [5] in 1998, and Conditional Random Fields (CRF) proposed by Mccallum A [5] in 2003. These traditional machine learning methods are usually composed of a system that reads large annotated corpus and uses a trained model to decode the sequence of test corpus, etc. However, traditional machine learning has the disadvantage of requiring high requirements for text extraction features. In recent years, the deep learning based NER methods got a fast development, in which the Neural Network method inputs the pre-trained word embedding to the Convolutional Neural Network (Convolutional Neural Network, CNN), Recurrent Neural Network (Recurrent Neural Network, RNN) etc., using large unmarked corpus word vector training, implements the end-to-end named entity recognition training.

Word2Vec is a widely used word embedding training tool in the field of natural language processing at present. However, Word2Vec is limited to its own training window size and can only deals with the words themselves, which does not conform to the different meanings or referential meanings displayed by different words according to the context in the actual language environment. Therefore, in the follow-up study, the application of long-distance dependent LSTM language model improves the training effect to a certain degree, but the fundamental problem is that in our daily language, all the information of words or phrases is understood context-dependent. However, for the above language training models, they all use the above information to predict the following information in a single way, or use the following information to predict the above information, which is difficult to meet the way of combining the context when human uses language, and the accuracy rate is not greatly improved.

Peters et al. [8] proposed Embedding from Language Models (ELMO) in 2018, which solves the problem that Word2Vec can only process word vectors statically, and to a certain extent solves the problem of single learning information. ELMO is essentially based on the process of dynamically adjusting the Word Embedding of each word in the current context.

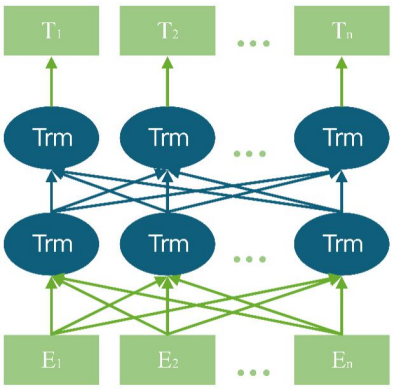

Figure 1: BERT model structure diagram.

The BERT model used in this paper has been widely used in experiments on language structure and language similarity in recent years [9–13], and has achieved good results. The overall architecture of the model is shown in Figure 1. The model uses a two-way Transformer, which absorbs the advantages of the model introduced above. The experimental results show that the two-way training language model has a deeper understanding of the context than the one-way language model.

3 INTRODUCTION TO PRE-TRAINING TASK OF BERT

BERT is a pre-trained language representation model for natural language processing, which is characterized by determining the main features in a text by calculating the weight of context relationship, and determining the real context relationship based on the fusion of context at all levels. Since in named entity recognition of Legal Case text, the semantic relationship between the identified object and the sentence is of great importance, the BERT Model pre-training includes two tasks which are Masked Language Model (Masked LM) and Next sentence prediction.

3.1 Masked Language Model

In the BERT Model, we adopted a mode of 'arbitrary masking' to complete the Deep bidirectional Representation task of training. During this training process, we only predict Masked Token which is called Masked Language Model (Masked LM).

The Masked Language Model (Masked LM) is descended from CBOW in Word2Vec. In Word2Vec, CBOW is similar to Masked LM, where CBOW predicts the central word based on what's around the central word, and does an earlier prediction for each of those words. Masked tags typically cover up to 15% of the sequence by default during training, which differs from the left-to-right language model pre-training in that Masked tags aim to predict Masked words based on context. The bi-directional transformer does not know which words are covered, so it will need to deal with the context of each word during the training process, randomly covering 15% of words similar to clothes-filling in English exams, without affecting the model's understanding of the overall language paragraphs.

Table 1: Examples for Next Sentence Prediction

Input Sentence	Label
[CLS] To promote the [MASK] achievements of our country's anti-drug work [SEP] Use [MASK] to commit crimes and educate the masses [SEP]	IsNext
[CLS] To promote the [MASK] achievements of our country's anti-drug work [SEP] and tried my best to [MASK] to excuse guilt [SEP]	NotNext
[CLS] The defendant Xu [MASK] fisted the victim Sun on the face [SEP] Caused [MASK] Nasal Bone Fracture [SEP]	IsNext
[CLS] The defendant Xu [MASK] fisted the victim Sun on the face [SEP] And truthfully confess [MASK] his own crime	NotNext

3.2 Next Sentence Prediction

Entity recognition task in Legal Case text not only needs to analysis the context relations between words, understanding and reasoning the relationship between the sentence is also needed in some important tasks while the relationship among the sentences cannot be directly modeled by the language model. Therefore, a binary task was adopted in BERT with the name of Next Sentence prediction which can be found in Table 1

Half part of the sentences in the input sequence is context-related sentences, and the other half part is randomly selected from the text. And we can used the Transformer model to judge the statement pairs in the text and determine whether there is a continuous relationship between the context sentences, so as to realize the modeling of the relationship between the statements.

At the same time, it also added some markers of specific action in the followings: the [CLS] marker is put in the first place of the first sentence, and the representation embedding obtained by BERT can be used for subsequent classification tasks. The [SEP] flag is used to separate two input sentences, such as input sentences A and B, and the [SEP] flag is added after the sentences A and B. The [MASK] flag is used to mask some words in the sentence, and after masking the words with the [MASK], the [MASK] embedding that BERT prints out predicts what the word is.

4 INTRODUCTION TO BERT-BILSTM-CRF MODEL

In the BERT-BiLSTM-CRF model, BERT is used as the feature representation layer to obtain word vectors, and then through BiLSTM deep learning full-text feature information, specific legal case entities are identified, and finally the CRF layer processes the BiLSTM output sequence and combines it in CRF The state transition matrix of, obtains a global optimal sequence according to the adjacent labels [14], and its structure can be found in Figure 2

4.1 Bi-LSTM Model

The LSTM model is an improved model proposed by Hochreiter S et al. [15] in 1997 for the problem of gradient disappearance and gradient explosion of RNN. BiLSTM is the abbreviation of Bidirectional Long Short-Term Memory, which is a combination of forward LSTM and backward LSTM. Usually the LSTM network encodes sentences from front to back and only masters the context information from front to back, and does not master the context

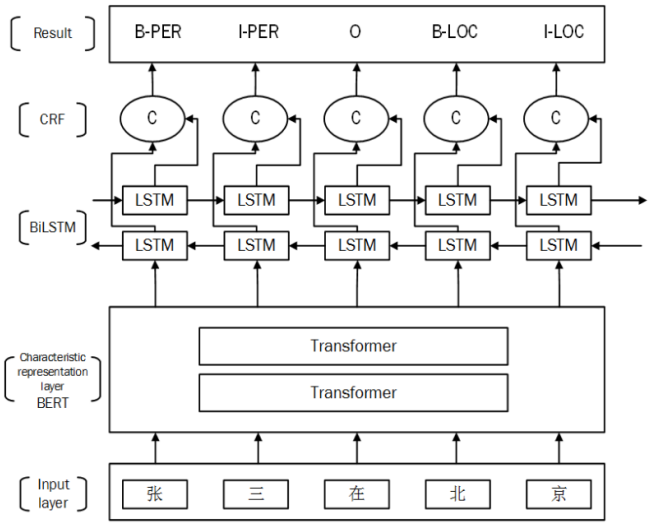

Figure 2: BERT model structure diagram.

information from back to front, so the forward LSTM network and the backward LSTM network are formed into a BiLSTM network to learn two-way context information [16]. The calculation process of LSTM can be summarized as follows: by forgetting the information in the unit state and storing new information, transmitting information useful for subsequent calculations, discarding unnecessary information, and outputting the hidden layer state at each time step. The state of the hidden layer at the previous moment and the forgetting gate, memory gate, and output gate calculated from the current input are used to jointly control the current forgetting, memory and output information. The overall process is shown in Figure 3

First of all, LSTM determines the information that needs to be discarded in the previous cell through the forget gate, and calculates a weight from 0 to 1 through formula (1) by receiving the output of the previous moment and the input of the current moment. And the weight means from completely being discarded to being completely reserved.

$$f_t = \sigma(W_f \cdot [h_{t-1}, X_t] + b_f) \tag{1}$$

The input gate controls the information that needs to be added to this Cell, and its calculation formula can be found in formula (2) and formula (3). The output gate is naturally used to control

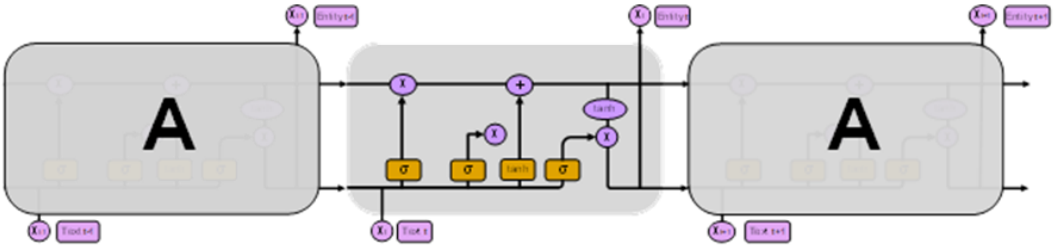

Figure 3: LSTM-Cell internal structure diagram.

which information is output as the task of the current stage. The calculation process is shown in formula (4) and formula (5). And Wi represents the weight matrix of the input gate, Wf represents the weight matrix of the forget gate, Wo represents the weight matrix of the output gate while b is the bias matrix of the three gates; while σ and tanh are the activation functions.

$$i_t = \sigma(W_i \cdot [h_{t-1}, X_t] + b_i) \tag{2}$$

$$C_t = f_t C_{t-1} + i_t \tanh(W_f \cdot [h_{t-1}, X_t] + b_c) \tag{3}$$

$$O_t = \sigma(W_0 \cdot [h_{t-1}, X_t] + b_0) \tag{4}$$

$$H_t = O_t \tanh(C_t) \tag{5}$$

4.2 CRF Model

The CRF is an algorithm that solves the conditional probability distribution of output random variables with specified random variables as input. In recent years, it has been widely used in the fields of part-of-speech tagging, syntactic analysis and named entity recognition [17]. CRF can consider the relationship between adjacent tagging results, combined with the existence of a large number of pronouns in the legal text, and actually obtain an optimal tag sequence result in the full text. The basic algorithm of CRF is as defined as followed:

$$S(x, y) = \sum_{i=0}^{n} A_{y_1, y_{i+1}} + \sum_{i=1}^{n} p_i \cdot y_i \tag{6}$$

$$P(y|x) = \frac{e^{S(x,y)}}{\sum_{\tilde{y} \in Y_x} e^{S(x,\tilde{y})}} \tag{7}$$

The output result of the BiLSTM layer is defined as Pnm, where n represents the number of words and m represents the tag category. Where Pij represents the probability that the i-th matches the j-th label. For the input sentence sequence x={x1,x2,...,xn} and its predicted sequence y={y1,y2,...,yn}, the probability expression is as shown in formula (6), all possible sequence paths After normalization, the probability distribution about the output sequence y is obtained, as shown in formula (7).

In the training process, the likelihood function of the label sequence, that is, to maximize the log probability of the correct label sequence y*, is shown in formula (8):

$$\log(p(y*|S)) = S(x, y*) - \log(\sum_{\tilde{y} \in Y_x} e^{S(x,\tilde{y})}) \tag{8}$$

In formula (8), Y_X represents the all possible mark sets, including sequences that do not comply with the BIO (beginning-inside-outside) ternary mark rule. The goal of adopting sentence-level likelihood function is to promote the model to generate correct tag sequences. Then the prediction result is a set of sequences with the largest overall probability output by equation 9):

$$y* = \arg\max K(x, \tilde{y}) \tag{9}$$

All the scores outputted by the BiLSTM layer will be used as the input of the CRF layer. The main feature of the CRF layer is that it can learn the implicit constraints of the sentence. For example, the first word in each sentence must be "B-" or "O", it is absolutely impossible to be "I-", because the beginning words of named entities such as "LOC", "ORG", and "PER" are all used "B-" means that there is absolutely no "I-" without "B-". In addition, a combination similar to "B-LOC I-ORG" must be wrong, because the adjacent "B-" and "I-" must represent the same type of entity. And the beginning of the named entity must be "B-", if there is a combination of "O I-", it must be wrong, and all entities start with "B-". With these basic principles, combined with the overall maximum probability calculated in the previous section, the wrong prediction sequence will be greatly reduced.

5 EXPERIMENTS AND ANALYSIS

5.1 Experimental Dataset and Data Indicators

This article used the corpus of the first half year in 1998 released by the Institute of Computational Linguistics of Peking University, the relevant Legal Case Texts of the CAIL Law Research Cups in year 2018 and 2019, the five types of case records with manually labelled entities by us, and other relevant texts of Legal Cases published on the Internet as the data set. The corpus of newspaper of People's Daily has been marked according to the ternary mark {B, I, O}, while B represents the first word of the classification entity, I represents the second word and the following words of the classification entity, and O represents the word that does not belong to a specific entity. At the same time, the dataset from the CAIL Law Research Cup and the relevant Legal Case Texts of the five types of case records are also marked by us. In five types of case records, LOC indicates the location of the case, ORG indicates the name of the specific organization, PER indicates the person in the case, TIME indicates the time involved in the case, MON indicates the money involved in the case, INJ indicates the injury involved in the case, and CRI indicates the name of the crime involved in the case.

This paper adopted the evaluation indicators of named entity recognition proposed in the MUC evaluation meeting. For each specific type of entity, accuracy rate (Precision), recall rate (Recall) and harmonic average (FB1) are used as evaluation criteria of model

Table 2: Definition for Experimental Evaluation Indicator

	The actual value is the relevant entity of the legal case	The actual value is not the relevant entity of the legal case
Test value is related to legal case	$A_{correct}$	$A_{incorrect}$
Test value is not related to legal case	A_c	A_d

Table 3: Algorithm Extraction Results of {B, I, O} triple

	Precision	Recall	FB_1
PER	84.23%	83.04%	92.24%
TIME	92.41%	92.06%	89.24%
LOC	88.31%	90.17%	83.42%
ORG	84.71%	82.13%	93.05%
MON	93.05%	93.05%	98.51%
INJ	97.01%	100%	90.51%
CRI	93.43%	87.59%	90.09%
AVERAGE	90.45%	89.72%	92.24%

performance which are defined as follows:

$$Precisison = \frac{A_{correct}}{A_{correct} + A_{incorrect}} \times 100\% \tag{10}$$

$$Recall = \frac{A_{correct}}{A_c + A_d} \tag{11}$$

$$FB_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{12}$$

5.2 Experimental Results and Analysis

Results in Table 3 show the extraction results of the {B, I, O} triple of BERT + BiLSTM + CRF model defined in the previous section. For the reason that there are much more abbreviations for specific organizations in the dataset, labelling result of the ORG organizations has a low accuracy rate relatively. And to our surprise, the accuracy of labelling the people characters in these Legal Cases of our model is particularly high while it has surpassed 97%.

The experimental results in Table 4 show that the experimental results of Word2Vec are not ideal. Although it is a word vector training tool widely used in the field of natural language processing, it has a small training window and only processes the words themselves. The problem does not meet the actual requirements for the difference of the same words in different contexts.

The experimental results of the two combination models of Word2Vec to obtain word vectors in terms of characters, locations, and organizational structure are somewhat different from those of the BERT word vector model. Its essence is a static solution. It is a high-complexity and pronouns refer to characters. And location recognition performance is not good.

Using the LSTM network to process the Word2Vec word vector results can be optimized for the need to increase the training window, and combined with the improvement requirements of the context as much as possible. In the experiment, a long-distance dependent two-way LSTM language model, namely BiLSTM, is used. The BiLSTM network can no longer only process a certain word itself, but integrate the long-distance context to obtain the maximum output score of the word itself, which is obviously greatly improved. However, since the output sequence of BiLSTM is based on the maximum value of the current word score, it is found in experiments that it is easy to subdivide the words that do not need to be subdivided, or there is often a problem of not considering the logical constraints of adjacent labels. To solve this problem, add the CRF layer. Through the comparison of the results of Word2Vec+BiLSTM and Word2Vec+BiLSTM+CRF recognition experiments, it can be found that the CRF model helps to consider the logic and order between the marked entity tags, and try to get the global optimal Sequence to further improve accuracy.

This paper proposes to use BERT instead of Word2Vec as a new word vector model. The biggest advantage of this layer is that it considers the reality of different semantics of the same word in different contexts. For example, there are the most "it", "plaintiff", "defendant" Pronouns that appear repeatedly in specific contexts such as ”basically refer to different content in different contexts, and have good effects in distinguishing and recognizing similar entities such as person names “Li Moujia” and “Li Mouyi”. Therefore, in the experimental results of BERT+BiLSTM+CRF, it can be seen that using BERT word vector training results to import BiLSTM+CRF has a greater improvement in the recognition accuracy of people, places and organizations than Word2Vec+BiLSTM+CRF. It shows that the BERT pre-training language model can be more accurate which can reflect the actual text information in the generated word vector.

6 CONCLUSION

Basing on the characteristics of Chinese Legal Case Texts, this paper has designed and proposed a BiLSTM+CRF based legal case entity identification method by using BERT as pre-training. In the usage of BERT language model, the most accurate semantics or references of words in the context can be analyzed to the most extent. At the

Table 4: Algorithm Result Comparison Table

	Word2Vec-BiLSTM-CRF			Word2Vec-BiLSTM			BERT-BiLSTM-CRF		
	Precision	Recall	FB_1	Precision	Recall	FB_1	Precision	Recall	FB_1
PER	84.23%	83.04%	83.64%	81.00%	78.42%	79.71%	97.94%	97.89%	97.92%
TIME	92.41%	92.06%	92.24%	90.61%	89.53%	90.07%	91.34%	90.16%	90.75%
LOC	88.31%	90.17%	89.24%	85.46%	86.14%	85.8%	95.48%	95.44%	95.46%
ORG	84.71%	82.13%	83.42%	81.04%	83.31%	82.18%	90.82%	92.91%	91.87%
MON	93.05%	93.05%	93.05%	92.66%	87.64%	90.15%	93.05%	94.46%	93.76%
INJ	97.01%	100%	98.51%	97.01%	100%	98.51%	100%	100%	100%
CRI	93.43%	87.59%	90.51%	92.70%	86.13%	89.42%	93.43%	96.31%	94.87%
AVERAGE	90.45%	89.72%	90.09%	88.64%	87.31%	87.97%	94.58%	95.31%	94.95%

same time, the CRF machine learning method is used to obtain a more accurate labeling sequence according to specific conditions. And according to the characteristics of domestic Chinese Legal Case Texts and using CRF to restrict the reality logic existing in legal intelligent entity recognition, it further improves the success rate of legal text entity recognition greatly.

After experimental evaluation, the three parameter values of Precision, Recall, and FB_1 of our method are all around 95%, reaching the leading level of the current existing public method. In this paper, only the places and persons in the Legal Case Text are marked accordingly at present and we will promote the model method to make it mark other elements of the Chinese Legal Case Text, such as time and case results. At the same time, we are doing our best to improve the accuracy of marking which can provide efficient solutions for the application of artificial intelligence technology into the Chinese judicial field for Legal Case Text processing.

ACKNOWLEDGMENTS

Thanks to the support of National key research and development program Project of China (No. 2018YFC0831301)

REFERENCES

[1] Vaswani A, N Shazzer, and N Parmar. 2017. Attention is all you need. *Advances in Neural Information Processing System.* (2017), 5998-6008. DOI https://dl.acm.org/doi/10.5555/3295222.3295349

[2] Grishman R, B Sundheim. 1996. Message understanding conference-6: A brief history. *Copenhagen, Proceedings of the 16th International Conference on Computational Linguistics (COLING 96), August 1996:* (1996),466-471. DOI https://dl.acm.org/doi/10.3115/992628.992709

[3] D M Bikel, R Schwartz, R M Weischedel.1999. An algorithm that learns what's in a name. *Machine learning.* (1999), 34(1): 211-231. DOI:https://dl.acm.org/doi/10.1023/A:1007558221122

[4] S Sekine, R Grishman, and H Shinnou. 1998. A Decision Tree Method for Finding and Classifying Names in Japanese Texts. *Money* (1998), 22: 112-220.

[5] A BORTHWICK. 1999. A Maximum Entropy Approach to Named Entity Recognitiont. *Ph. D. Thesis New York University* . (1999):4701-4708.

[6] A MCCALLUM. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. *Proc. CoNLL* (2013), 188–191. DOI: https://dl.acm.org/doi/10.1109/NLPKE.2005.1598798

[7] Y Zhang, J Yang. 2018. Chinese NER Using Lattice LSTM. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* .(Volume 1: Long Papers). (2018): 1554-1564. DOI: https://dl.acm.org/doi/10.18653/v1/P18-1144

[8] M E Peters, M Neumann, M Iyyer. 2018. Deep contextualized word representations. *Proceedings of NAACL-HLT.* (2018), 2227–22376. DOI: https://dl.acm.org/doi/10.18653/v1/N18-1202

[9] G Jawahar, B Sagot, D Seddah. 2019. What Does BERT Learn about the Structure of Language?. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* (2019), 3651–3657. DOI: https://dl.acm.org/doi/10.18653/v1/P19-1356

[10] N Peinelt, D Nguyen, M Liakata. 2020. tBERT: Topic models and BERT joining forces for semantic similaritydetection. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* (2020), 7047–7055. DOI:https://dl.acm.org/doi/10.18653/v1/2020.acl-main.630

[11] D Chai, W Wu, and Q Han. 2020. Description based text classification with reinforcement learning. *International Conference on Machine Learning. arXiv:* 2002.03067

[12] C Qu, L Yang, M Qiu. 2019. BERT with history answer embedding for conversational question answering. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval,* (2019): 1133-1136. DOI:https://dl.acm.org/doi/10.1145/3331184.3331341

[13] Z Dai, J Callan. 2019. Deeper text understanding for IR with contextual neural language modeling. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval.* (2019), 985–988. DOI: https://dl.acm.org/doi/10.1145/3331184.3331303

[14] Zhao Ping, Lianying, Sun, Ying Wan. 2020. Named Entity Recognition of Chinese Scenic Spots Based on Bert + Bilstm +CRF. *Computer Systems and Applicationsc.* (2020), 29(06): 169–174. DOI: https://dl.acm.org/doi/10.15888/j.cnki.csa.007269

[15] S Hochreiter, J Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation,* (1997) , 9(8): 1735-1780. DOI: https://dl.acm.org/doi/10.1162/neco.1997.9.8.1735

[16] Yushuai Zhang, Huan Zhao, and Bo Li. 2021. Semantic Slot Filling Based on Bert and Bilstm. *Computer Scienceroc.* (2021), 48(1): 247-252. DOI: https://dl.acm.org/doi/10.1162/neco.1997.9.8.1735

[17] Lishuang Li, Yuankai Guo. 2018. Biomedical Named Entity Recognition Based on CNN-BLSTM-CRF Model. *Journal of Chinese Information Processing.* (2018), 32(1): 116-122. DOI:https://dl.acm.org/doi/CNKI:SUN:MESS.0.2018-01-016.