A Method for Resume Information Extraction Using BERT-BiLSTM-CRF

XiaoWei Li
State Key Laboratory of MEAC
Information Engineering University
Zhengzhou, China
shelwei@163.com

Hui Shu
State Key Laboratory of MEAC
Information Engineering University
Zhengzhou, China
415314938@qq.com

Yi Zhai
State Key Laboratory of MEAC
Information Engineering University
Zhengzhou, China
amber_pro@163.com

ZhiQiang Lin
State Key Laboratory of MEAC
Information Engineering University
Zhengzhou, China
linzq1064@163.com

2021 IEEE 21st International Conference on Communication Technology (ICCT) | 978-1-6654-3206-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCT52962.2021.9657937

***Abstract*—To solve the problem of low efficiency of electronic resume information extraction by artificial construction rules, a resume information extraction method based on named entity recognition is proposed, which extracted personal details such as graduation college, job intention and job skills from the resume into named entity recognition. Firstly, the TXT text in different formats of resume file is extracted for data cleaning and other preprocessing. The BERT language model based on multi-head self-attention mechanism is used to extract text features and obtain word granularity vector matrix. The BiLSTM neural network is used to obtain the context abstraction features of serialized text. Finally, using CRF to decode and annotate the global optimal sequence, the corresponding resume entity information is extracted. Experimental results show that the whole scheme can effectively extract electronic resume information, and the performance of the resume information extraction model based on BERT-BiLSTM-CRF is better than other models.**

Keywords—resume parsing, information extraction, named entity recognition, BERT

I. INTRODUCTION

With the rapid development of the Internet, the total amount of various media data has exploded. A resume is an official document for job seekers to show their work experience and skills to online recruitment websites or company human resources. Recruitment portals on the Internet receive a large number of personal resumes from job applicants. These massive personal data have great research value.

Information extraction is a text processing technology that extracts entities, relationships, events and other information of specified types from a large number of natural language texts and forms structured data output. At present, information extraction has become an important research issue in the field of natural language processing. Resume information extraction refers to extracting information interested by employers from resumes in different formats and forms for structured storage, including basic personal information, education experience, work experience, professional skills, interests and hobbies, etc. [1]. Resume information extraction realizes automatic filing and management of resume information and provides information sources for subsequent resume retrieval, industry classification and talent recommendation [2].

Resume information extraction can be achieved by using the named entity recognition task in natural language processing. Named entity recognition is usually regarded as a sequence labeling task to identify the names of people, places, organizations and other information in the text. Named entity recognition based on machine learning mainly uses hidden Markov model (HMM) and conditional random field (CRF) model. In recent years, thanks to the development of word vector technology and deep learning, the named entity recognition model based on the neural network model has achieved better results [3]. Existing resume information extraction model based on neural network models usually uses static word vector methods such as Word2Vec to encode resume text, and then send it to convolutional neural network (CNN) or recurrent neural network (RNN) for training, and then uses the trained model to extract resume information [4]. The existing resume extraction model using Word2Vec, a context-free word vector tool, has weak text representation ability and cannot solve the problem of polysemism. In this paper, the dynamic pre-training model BERT with stronger feature representation ability is used as the feature representation layer is combined with the bidirectional long and short-term memory network (BiLSTM) and conditional random field to construct an information extraction model. The experimental verification is carried out using the resume data set that has undergone label format conversion. The results show that the BERT-BiLSTM-CRF model helps to improve the accuracy of resume information extraction.

II. RELATED WORKS

Early research on resume information extraction based on machine learning was mainly based on HMM, CRF, Support Vector Machine (SVM), etc. to build an information extraction model [5-7]. These models reduce the cost of manual maintenance of rules, but a lot of data needs to be prepared. The accuracy is related to the quality and quantity of data selection to a certain extent. When faced with the problem of complex data generalization, the ability is limited [8].

In recent years, deep learning methods have been used to extract resume information. In 2018, Huang et al. [9] proposed a resume information entity extraction method based on LSTM and CRF joint model. The method firstly initializes the input word sequence through Word2Vec training, then fuses the context information of the words to be labeled with the

978-1-6654-3206-1/21/$31.00 ©2021 IEEE

bidirectional LSTM layer, outputs the scores of all possible tag sequences to the CRF layer, and finally obtains the optimal tag sequence by the constraints between the tags before and after their introduction. Pham et al. [10] established an information extraction model for resumes in DOC, DOCX, PDF, TXT and other formats by using the rule-based method combined with deep learning. Comparative experimental results showed that CNN-BI-LSTM-CRF model was significantly superior to other models. It shows that character-level representation plays an important role in language sequence annotation. Katsuta et al. [11] formalized resume information extraction into a sequence labeling problem, annotating English and Japanese resume corpora, and assisted by the third-party tool NeuroNER [12] for experiments. Ayishathahira et al. [13] proposed a resume parsing system combining convolutional neural network (CNN), bi-directional long short-term memory (BI-LSTM), and conditional random field (CRF) deep learning models. CNN model is used to classify different text blocks in resumes, and CRF and Bi-LSTM-CNN models are used for named entity recognition. In 2019, Chen et al. [14] proposed a feature fusion based Chinese resume parsing method, that is, cascading word vectors generated by Word2Vec and word vectors generated by bi-LSTM modeling word sequences, and then combining BI-LSTM and CRF to parse Chinese resumes. In 2020, Zu et al. [15] proposed an end-to-end pipeline for RESUME parsing based on neural network text classifier and word vector. The resume parsing pipeline can combine upstream text block segmentation with downstream specific information recognition. In concrete information extraction, various sequence annotation classifiers identify named entities in segmented text blocks. Compared with the recognition performance of the four sequence annotation classifiers, the advantages of BLSTM-CNNS-CRF in named entity recognition are established.

Different from the above work, we first use dynamic pre-training model BERT instead of Word2Vec to express the words embedded in the resume, and then establish the resume information extraction model combined with neural network, so that the extraction model can extract the deep features of the resume text and achieve more accurate resume information extraction.

III. Resume Information Extraction Scheme

The overall framework of resume information extraction proposed in this paper is shown in Figure 1. Job applicants submit electronic resumes in various formats, such as PDF, DOC, DOCX formats, etc. Therefore, before extracting information on the resume, it is necessary to grab the plain text information from different formats of electronic resumes; then perform data cleaning on the resume text. For example, removing special symbols in the text, unifying punctuation, etc. Considering that the basic personal information in the resume, such as email address, phone number, and other attributes have obvious grammatical characteristics, we directly use regular expression-based rules to extract such simple information. For the difficult to extract personal details such as educational background, job-seeking intention, work skills, work experience, etc., which are not obvious grammatical features, we construct a named entity recognition model based on BERT-BiLSTM-CRF to extract information from it.

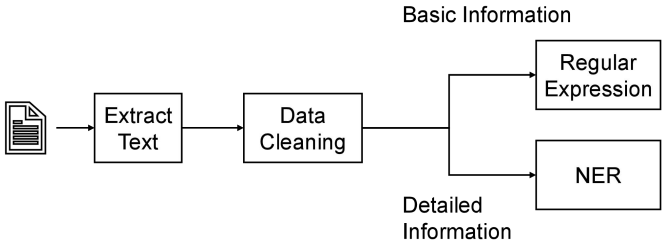

Fig. 1. The overall framework of resume information extraction.

Pdfminner, docx2txt, textract, doctotext, pdftotext and other tools were used to extract different formats of resume files. Finally, it is determined that the use of doctotext and pdftotext tools for PDF, DOC, DOCX text extraction efficiency is the highest. After obtaining the resume text file, the data of the resume text is cleaned, and the unprintable characters are removed. It can be observed that the basic information of the name, email and telephone number in the resume has obvious grammatical and lexical characteristics, which can achieve very high accuracy by constructing rules to extract such information. E-mail in the personal basic information, for example, the '@' and '.' these two characters, by using the regular expression [^ @ | \ s] + @ [^ @] + \. [^ @ | \ s] + for extraction. Rule-based resume information extraction is not the focus of this paper. Below we focus on the detailed resume information extraction based on the BERT-BiLSTM-CRF model.

IV. BERT-BiLSTM-CRF Model

The BERT-BiLSTM-CRF information extraction model framework constructed in this paper is shown in Figure 2, which consists of BERT module, BiLSTM module and CRF module [16,17].

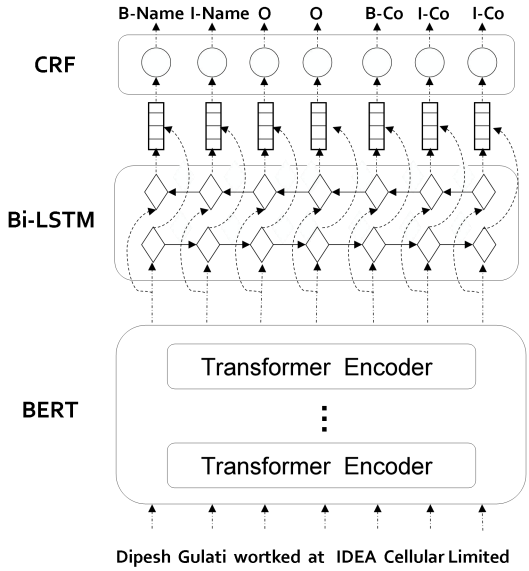

Fig. 2. The framework of BERT-BiLSTM-CRF model.

Firstly, the resume text after word segmentation is input into the pre-trained BERT model to obtain the word vector of the resume text. Then it is sent to BiILSTM to learn the contextual information of the resume. Finally, the output

sequence of BiLSTM is sent to CRF layer, and the CRF layer solve the global optimal sequence according to the state transition matrix and the labels between neighbors.

A. BERT

Before 2018, Word2Vec [18] was a word vector training tool widely used by scholars, which promoted the wide application of deep learning in natural language processing tasks. Word2Vec is a static method, the relationship between words and vectors is one-to-one, so it cannot solve the problem of ambiguity of a word, and the information it expresses has no deep contextual information. BERT [19] (Bidirectional Encoder Representation from Transformers) is a dynamic pre-trained language model, which adopts bi-directional Transformer technology, and its feature Representation depends on the left and right contexts. The dynamic word vector can be trained more accurately, which leads to the explosive development of natural language processing. At present, BERT is more and more widely used in natural language processing, including text classification, reading comprehension, information extraction, automatic papers, automatic question and answer, etc.

BERT is the first part of the BERT-BiLSTM-CRF model. It is mainly responsible for converting the original text of the input resume into a vector form, and then sending vectors to BiLSTM to continue the learning of context features. BERT is a dynamic pre-training model. It has two tasks during training. One is to predict the words that are randomly masked in the sentence, and the other is to determine whether the input two sentences are the upper and lower sentences in the paper. In the first task, randomly mask 15% of the words in each sentence, 80% of the words are replaced by the token [Mask], 10% of the words are replaced by random words, and 10% of the words remain unchanged, then train the model to predict the words that are masked. In the task of predicting the next sentence, 50% of the sentences in the corpus select their corresponding next sentence together to form a sentence pair as a positive sample; the remaining 50% sentences randomly select a non-next sentence as a negative sample for training.

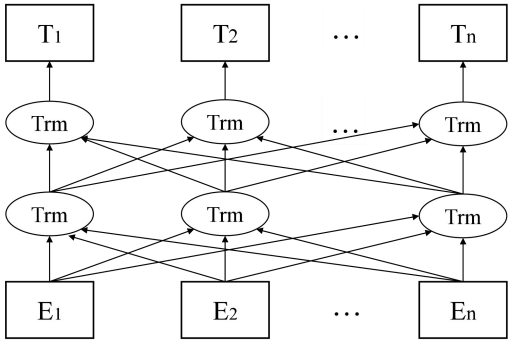

Fig. 3. The composition structure of the BERTmodel.

The main structure of BERT is shown in Figure 3, where Trm represents the encoder based on the self-attention mechanism, which is the encoder part of the standard Transformer structure; E1, E2, ⋯, En represent the input of the model; T1, T2, ⋯, Tn represents the output of the model. Each layer of BERT is composed of a Trm unit. In the basic version of the BERT model, there are a total of 12 layers of Trm, each layer has 12 Attentions, and the word vector dimension is 768. General language models have limited understanding of the relationship between sentences, and the semantic relationship between sentences is very important for named entity recognition. The BERT model captures the deep semantic relationship between sentences and can solve the ambiguity problem of a word, thereby improving the performance of the named entity recognition task. Therefore, this paper combines the BERT language model into the task of resume information extraction and has achieved significant results.

B. BiLSTM

BiLSTM is the second part of the BERT-BiLSTM-CRF model. It can further capture contextual features and obtain more comprehensive semantic information of the resume text. BiLSTM is a combination of forward LSTM and backward LSTM. Long-short-term memory network (LSTM) is a special recurrent neural network [20], which can learn the long-term dependence information of the sequence, and solve the problem of gradient disappearance and gradient explosion in the training process of long sequence. LSTM is mainly composed of forget gate, input gate and output gate, and its typical structure is shown in Figure 4.

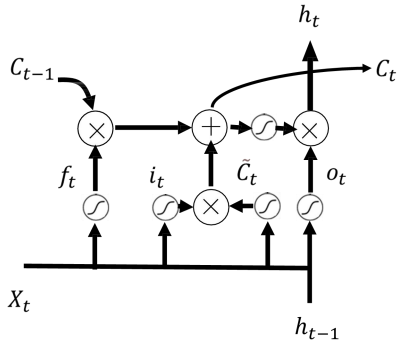

Fig. 4. The typical structure of the BiLSTM model.

- Forget gate: This control unit mainly selectively forgets the input information of the previous unit node. The calculation formula is shown in Equation (1): where f_t is the output value of the forget gate, W_f is the weight matrix, h_{t-1} is the state of the hidden layer at the previous moment, X_t is the input at the current moment, and b_f is the bias.

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

- Input gate: This control unit mainly retains the input information of the current unit node selectively. The calculation formulas are shown in Equations (2), (3), and (4): where i_t is the output value of the input gate, $\tilde{C}_t$ represents the temporary state of the current unit node, and C_t represents the state of the current unit node.

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

$$\tilde{C}_t = tanh\left(W_C \cdot \left[h_{t-1}, X_t\right] + b_C\right) \quad (3)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (4)$$

- Output gate: This control unit mainly selectively outputs the information of the current time node. The calculation formulas are shown in Equations (5) and (6), where o_t is the output value of the output gate, and h_t is the state of the hidden layer at the current moment.

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, X_t\right] + b_o\right) \quad (5)$$

$$h_t = o_t * tanh\left(C_t\right) \quad (6)$$

The network structure of LSTM can effectively filter and memorize the information of the memory unit through training, and capture the long-distance dependence. However, there is still a problem with LSTM to model resume text: it can only obtain the information from the previous text, and cannot encode the information from the back to the front. In the task of extracting resume information, the front and back text information helps to better identify named entities. Therefore, we choose the BiLSTM model, using the word vector output by BERT as the input of BiLSTM, and the forward and backward LSTM are respectively used to obtain the information hidden before and after the resume text, and then the two information is combined and sent to CRF for the identification of the resume information entity.

C. CRF

Conditional random field (CRF) is the last part of BERT-BiLSTM-CRF model, which is responsible for capturing the dependencies between the previous and subsequent tags corresponding to the resume text, and better predicting the resume entity tags by using the global information of the tag sequence. Conditional random field is a discriminant probabilistic undirected graph model, which is usually used to analyze and label sequence data [21]. The model combines the advantages of maximum entropy model and hidden Markov model. In recent years, it has achieved good results in sequence tagging tasks such as word segmentation, part of speech tagging and named entity recognition.

The linear chain conditional random field is used in this paper, which refers to the conditional probability distribution model satisfying the Markov property when the observation sequence *X* and the state sequence *Y* are represented by linear chains. Where: the observation sequence *X* corresponds to the resume text sequence in this paper, and the output sequence *Y* is the entity label category corresponding to the resume text sequence. The parametric form of linear chain conditional random field is shown in equations (7) and (8):

$$P\left(Y|X\right) = \frac{1}{\mathrm{Z}\left(X\right)} exp \sum_{k=1}^{K} \omega_k f_k\left(Y, X\right) \quad (7)$$

$$Z\left(X\right) = \sum_{Y} exp \sum_{k=1}^{K} \omega_k f_k\left(Y, X\right) \quad (8)$$

where: $Z(X)$ is the normalization factor; f_k is the characteristic function, ω_k represents the weight corresponding to the characteristic function. During model training, the resume text training set is used to obtain the conditional probability model $P(Y|X)$ through maximum likelihood estimation. In the resume text prediction, for a given resume text sequence, the dynamic programming algorithm Viterbi is used to solve the optimal output sequence that maximizes the $P(Y|X)$, so as to obtain the entity label information corresponding to the resume text.

V. Experiments and Analysis

In this paper, the above resume information extraction model based on BERT-BiLSTM-CRF is implemented, and the validity of the model is evaluated by experiments and analysis using English resume data set.

A. Dataset

In order to extract information from resumes, the annotated resume corpus is needed to train and test the information extraction model. Part of the resume data set in this study is from the annotated resumes provided by Dataturks official annotation project[1]. A total of 700 annotated English resumes are collected and received in this paper. The label types marked on the dataset include Name, Designation, Location, Skills, College Name, Degree, Companies worked at, Years of experience, etc. The annotation format of the resume dataset is Dataturks format, which cannot be directly used for training the model in this paper, so we convert the annotation format of the resume dataset. We adopt the BIO annotation format (B-begin, I-inside, O-outside). An example of annotation is shown in Table I:

TABLE I. The Examples of Corpus Labeling

Token	Harry	Potte	Graduated	from	Hogwarts	School
Lable	B-Nam	I-Nam	O	O	B- Col	I- Col

B. Evaluation Metrics

This paper selects three metrics: precision, recall and F1 value (f1-score) to evaluate the performance of the model. The definitions are shown in Table II. Among them, TP is the number of information entities that can be correctly recognized by the model, FP is the number of irrelevant information entities that are recognized by the model, and FN is the number of related information entities that are not recognized by the model.

TABLE II. Definition of Evaluation Metrics

Metrics	**Definition**
Precision	$P = \frac{TP}{TP + FP}$
Recall	$R = \frac{TP}{TP + FN}$
F1-score	$f1 = \frac{2PR}{P + R}$

C. Experimental Setups

Python3.7 and TensorFlow are used to build the model in 64-bit Ubuntu 20.04 operating system. The computer is

[1]https://github.com/DataTurks-Engg/Entity-Recognition-In-Resumes-SpaCy

equipped with an 8-core 16-thread CPU, 16GB of ram, and an NVIDIA GeForce RTX 2080 graphics card. During model training and evaluation, the sequence length was 512, batch size was 4, epoch was 100, the optimizer was Adam, the initial learning rate was 0.001, and the number ratio of training set and test set was 4:1.

D. Results Evaluation

In this paper, the collected resume data set is cleaned including removing line breaks and some special interference characters, then the resume data set annotation format is converted to BIO format, and then the data is sent into the model for training. In order to compare and evaluate the performance of BERT-BilSTM-CRF extraction model, this paper uses resume data set to conduct comparative experiments on the following four models, and observe the evaluation metric results of the experiment.

a) BiLSTM: the pre-trained Word2Vec word vector is used, and then send them to BiLSTM for resume entity annotation.

b) BiLSTM-CRF: The pre-trained Word2Vec word vector was used, and then send them to BiLSTM-CRF for resume entity annotation.

c) BERT-BiLSTM: The word vector output by the BERT model is sent to BiLSTM for resume entity annotation.

d) BERT-BiLSTM-CRF: The resume information extraction model proposed in this paper.

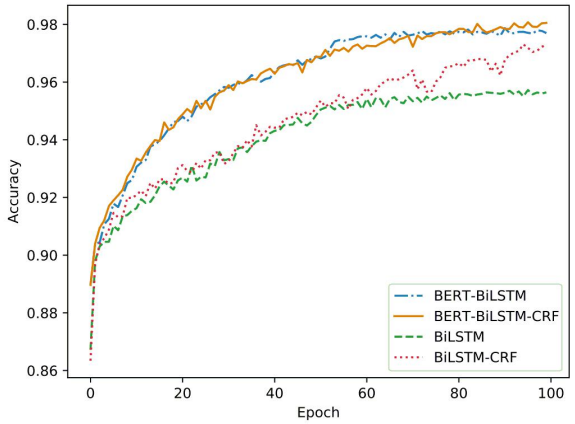

Fig. 5. The training property of each model.

Figure 5 shows the training of the four models in 100 epochs. It can be seen from the figure that the two models that use the BERT model for word embedding have better effects than the two models that use Word2Vec for word embedding. In the two models that both use BERT for word embedding, the performance of the BERT-BiLSTM-CRF model is better than that of the BERT-BiLSTM model, indicating that adding CRF can further improve the performance of the model.

Table III shows the recognition results of the main information entities of the BERT-BiLSTM-CRF model on the resume test set. It can be seen that the recognition accuracy of the Name entity is the highest, with an F1 value of 97.64%, and the recognition accuracy of the Designation entity is the lowest. The F1 value is 86.43%.

TABLE III. BERT-BiLSTM-CRF EXPERIMENTAL RESULTS (%)

Label	Precision	Recall	F1-score
Name	97.32	97.97	97.64
Designation	88.06	84.69	86.43
Location	92.36	91.10	91.72
Skills	95.28	87.40	91.17
College Name	88.73	86.30	87.50
Degree	94.59	90.91	92.72
Companies worked at	88.65	84.63	86.60
Years of Experience	93.94	91.18	92.54

Table IV shows the evaluation metric results of the four models under the same experimental environmental conditions and training parameters. It can be seen from Table IV that the performance of the two models that use BERT for word embedding is better than the model that does not use BERT for word embedding. After 100 rounds of training, the precision rate of the BERT-BiLSTM-CRF model is 91.41%, the recall rate is 88.17%, and the F1 value is 89.69%. All indicators have reached the highest level. Comparative experiments show that when building a resume information extraction model, the BERT-BiLSTM-CRF model has stronger feature extraction capabilities and higher information extraction accuracy than other models.

TABLE IV. COMPARATIVE EXPERIMENTAL RESULTS (%)

	Precision	Recall	F1-score
BiLSTM	47.54	41.52	43.08
BiLSTM-CRF	73.85	56.28	62.02
BERT-BiLSTM	79.38	79.16	79.11
BERT-BiLSTM-CRF	**91.41**	**88.17**	**89.69**

VI. CONCLUSION

Aiming at the problem of low accuracy of resume information extraction and difficulty in extracting features by manual construction, we propose a resume information extraction method fused with the BERT language model. The contextualized word vector of the resume text is obtained through the BERT pre-training language model. Combined with the BiLSTM deep neural network to fully learn the text context information and the CRF machine learning model to calculate the global optimal annotation sequence, the Bert-BilSTM-CRF model was constructed. And we conducted a comparative experiment using the English resume dataset. The experimental results show that the proposed method in this paper can effectively extract resume information, and has better information extraction performance than other models. In the next step, the model can be further optimized to improve the training efficiency of the model. In the future, we plan to apply this model to other fields to perform information extraction tasks in corresponding fields.

REFERENCES

[1] Deshpande, Amala, et al. "Proposed system for resume analytics." Int. J. Eng. Res. Technol. (IJERT) 5.11 (2016): 468-471.

[2] Deepak, Gerard, Varun Teja, and A. Santhanavijayan. "A novel firefly driven scheme for resume parsing and matching based on entity linking paradigm." Journal of Discrete Mathematical Sciences and Cryptography 23.1 (2020): 157-165.

[3] Yadav, Vikas, and Steven Bethard. "A survey on recent advances in named entity recognition from deep learning models." arXiv preprint arXiv:1910.11470 (2019).

[4] Chou, Yi-Chi, and Han-Yen Yu. "Based on the application of AI technology in resume analysis and job recommendation." 2020 IEEE International Conference on Computational Electromagnetics (ICCEM). IEEE, 2020.

[5] Yu, Kun, Gang Guan, and Ming Zhou. "Resume information extraction with cascaded hybrid model." Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05). 2005.

[6] Pawar, Sachin, Rajiv Srivastava, and Girish Keshav Palshikar. "Automatic gazette creation for named entity recognition and application to resume processing." Proceedings of the 5th ACM COMPUTE Conference: Intelligent & scalable system technologies. 2012.

[7] Wentan, Yan, and Qiao Yupeng. "Chinese resume information extraction based on semi-structured text." 2017 36th Chinese Control Conference (CCC). IEEE, 2017.

[8] Mu, Ruihui. "A survey of recommender systems based on deep learning." Ieee Access 6 (2018): 69009-69022.

[9] Huang Sheng, et al. "Entity extraction method of resume information based on deep learning." Computer Engineering and Design, 2018, 39(12): 3873-3878.

[10] Pham Van, Long, Sang Vu Ngoc, and Vinh Nguyen Van. "Study of Information Extraction in Resume." Conference, 2018.

[11] Katsuta, Akihiro, et al. "Infomation extraction from english & japanese résumé with neural sequence labelling methods." (2018).

[12] Dernoncourt, Franck, Ji Young Lee, and Peter Szolovits. "NeuroNER: an easy-to-use program for named-entity recognition based on neural networks." arXiv preprint arXiv:1705.05487 (2017).

[13] Ayishathahira, C. H., C. Sreejith, and C. Raseek. "Combination of neural networks and conditional random fields for efficient resume parsing." 2018 International CET Conference on Control, Communication, and Computing (IC4). IEEE, 2018.

[14] Chen, Jiaze, Liangcai Gao, and Zhi Tang. "Information extraction from resume documents in pdf format." Electronic Imaging 2016.17 (2016): 1-8.

[15] Zu, Shicheng, and Xiulai Wang. "Resume information extraction with a novel text block segmentation algorithm." Int J Nat Lang Comput 8 (2019): 29-48.

[16] Dai, Zhenjin, et al. "Named entity recognition using bert bilstm crf for chinese electronic health records." 2019 12th international congress on image and signal processing, biomedical engineering and informatics (cisp-bmei). IEEE, 2019.

[17] Jiang, Shaohua, et al. "A BERT-BiLSTM-CRF model for Chinese electronic medical records named entity recognition." 2019 12th International Conference on Intelligent Computation Technology and Automation (ICICTA). IEEE, 2019.

[18] Rong, Xin. "word2vec parameter learning explained." arXiv preprint arXiv:1411.2738 (2014).

[19] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[20] Sundermeyer, Martin, Ralf Schlüter, and Hermann Ney. "LSTM neural networks for language modeling." Thirteenth annual conference of the international speech communication association. 2012.

[21] Lafferty, John, Andrew McCallum, and Fernando CN Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." (2001)