Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 183 (2021) 706–712

Procedia Computer Science

www.elsevier.com/locate/procedia

10th International Conference of Information and Communication Technology (ICICT-2020)

BiLSTM-Attention-CRF model for entity extraction in internet recruitment data

Xia Cui, Feifei Dai, Changpeng Sun, Zihua Cheng*,
Borang Li, Bo Li, Yaoxin Zhang, Zhongjun Ji, Deyu Liu

State Grid Tianjin Electric Power Company, Tianjin 300010, CHN

Abstract

With the development of the Internet, online recruitment has gradually become the mainstream. Analyzing the recruitment data and exploring the internal rules of the data can help enterprises and job seekers realize human-career matching. An extremely important step in analyzing recruitment data is to extract structured information from unstructured data. Named entity recognition can effectively extract entity information from unstructured data. In recent years, a lot of work has focused on this task, but no related work has been applied named entity recognition to Internet recruitment information. Therefore, this paper proposes the BiLSTM-Attention-CRF model for Internet recruitment information, which can be used to extract skill entities in job description information. This model introduces the BiLSTM and Attention mechanism to improve the effect of entity recognition. To verify the performance of the model, the paper used crawler technology to capture the real Internet recruitment data and annotate it and obtain a real data set. In this paper, a series of experiments are conducted on this data set, and the experimental results show that the proposed model achieves the best performance.

Keywords: Name Entity Recognition; Internet Recruitment; Attention Mechanism

1. Introduction

Talents play an important role in enterprises' development. Recruiting the right employees becomes a critical factor in every company's business strategy. Recently, online Internet recruitment is favored by employers and job

* Corresponding author. Tel.: +86-13820119882.
E-mail address: 1309184547@qq.com

10.1016/j.procs.2021.02.118

seekers due to its advantages such as rich information and strong real-time. The rapid progress of online recruitment has produced a large amount of recruitment data. By analyzing the data characteristic, employers can work out more reasonable information, and job seekers can get more matching job information.

However, Internet recruitment data is not entirely structured, and data analysis requires structured data. We need to extract useful fine-grained information from unstructured data like job descriptions, company details, and so on. Therefore, how to extract fine-grained information accurately and efficiently from recruitment data is an important research topic. Named Entity Recognition (NER) is one of the common tasks in natural language processing, effectively solving the above problems. Nevertheless, there is a lack of datasets of Internet recruitment information for related research.

Therefore, this paper constructs an Internet recruitment information dataset by collecting and manually labeling recruitment data. For the unstructured job description information in the data, this paper proposes a conditional random field model based on attention mechanism and bidirectional long short-term memory neural network (Bi-LSTM-ATTENTION-CRF, BAC for short) to extract the skill entities in the job description. the model obtains the word vector representations by Bert, and extracts semantic information using Bi-LSTM. Then CRF is used for sequence labeling to obtain the skill entities in the job description information. Considering the problem of long-distance dependency in the text, the model incorporates the attention mechanism to learn long-distance semantic features. The experimental results in this paper show that the model can effectively extract skill entities from job descriptions. Simultaneously, the proposed method is not limited to the recruitment field but can be widely applied to structure unstructured text in various fields, including audit, law, and education. Taking auditing as an example, unstructured texts greatly limit audit work quality and efficiency, including problem descriptions, institutional basis, and audit opinions.

2. Related work

NER is a fundamental part of many NLP tasks, such as machine translation[1], knowledge base construction[2], automatic question answering[3], and search[4].

Earlier, NER mainly utilized rules and dictionaries to extract entities. Quimbaya[5] proposed a dictionary-based approach for electronic health records. Kim[6] used manual rules to extract entities in speech. These methods depended heavily on the quality of the dictionary and the coverage of the rules.

To reduce manual construction of rules and dictionaries, researchers have proposed many end-to-end models. Li[7] exploited SVM to predict entity labels, but this method does not take adjacent words into account in the prediction procedure. Szarvas[8] introduced training as an independent decision tree classifier and using a voting scheme to determine the final decision. McCallum and Li[9] developed a CRF-based method and achieved good results on CoNLL03 datasets. As deep learning advances, neural networks are also applied in the NER field. Collobert[9] et al. combined Convolutional Neural Networks with CRF. Kuru[10] proposed CharNER to extract character-level features. Taiki[16] et al. applied NER to the area of compound description. Xiao[17] used auxiliary classifiers to tackle the problem of entity fragmentation in NER. Zihan[18] put forward a cross-weighting method to deal with labeling errors in the NER dataset. Liu[19] added a geographical dictionary to the neural network model to improve entity extraction.

While the above approaches have significantly improved the quality of entity extraction and facilitated the development of many downstream tasks, little work has been done in the area of Internet recruitment. Furthermore, existing models do not perform well in addressing long-distance dependency. We focus on the task of extracting entities from Internet job information and proposes a BAC model to extract the skill entities from job descriptions.

3. CRF based on attention mechanism and Bi-LSTM

Given job description $X = \{x_1, x_2, ..., x_n\}$, representing text containing n words. NER aims to extract a specific entity $T = \{t_1, t_2, ..., t_n\}$. Each entity is included in the given text, i.e. the entities are all subsets of X.

To better tackle the task of skill entity extraction, we propose the BAC model, the overall structure is shown in Figure 1. Specifically, we first use BERT for word embedding, then extract semantic information with bidirectional

LSTM. To strengthen the semantic relationship of distant information, we introduce the attention mechanism[12] to our model. Finally, the model adopts CRF[13] to obtain the sequence labeling results.

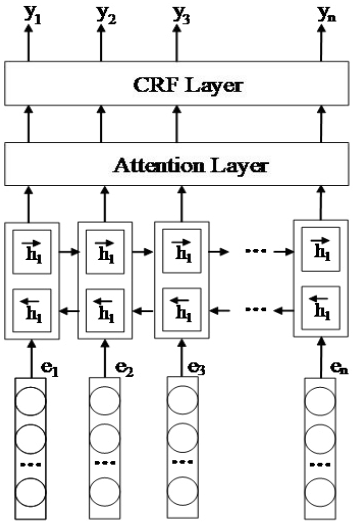

Fig. 1. BAC model.

3.1. Word embedding

Given input $X = \{x_1, x_2, ..., x_n\}$, where n means the number of words, the model maps X into a series of word vectors $E = \{e_1, e_2, ..., e_n\}$. We adopt BERT to generate representations, which is a pre-trained model that uses Masked LM and Next Sentence Prediction tasks to learn representations. The same word may express different meanings in different sentences, and BERT assigns different representations to words based on contextual information. Thus, the word vectors obtained by BERT enable better results for downstream tasks.

3.2. LSTM layer

Based on the previous process to obtain the word vectors $E = \{e_1, e_2, ..., e_n\}$ this paper uses LSTM[14] to model the contexts' semantic information and capture the semantic features. LSTM is a variant of RNN characterized by "forgetting gate" and "input gate", which can effectively alleviate long-term dependence problem. In detail, the calculation process of LSTM at the t-th time step is as follows:

$$\begin{aligned} &z_t = \tanh(W_z[h_{t-1}, e_t]), i_t = \text{sigmoid}(W_i[h_{t-1}, e_t]) \\ &f_t = \text{sigmoid}(W_f[h_{t-1}, e_t]), o_t = \text{sigmoid}(W_o[h_{t-1}, e_t]) \\ &c_t = f \odot c_{t-1} + i \odot z, h_t = o \odot \tanh c_t \end{aligned} \tag{1}$$

where f_t and o_t denote the input, forgetting and output gates of the LSTM, respectively. h_t is the t-th timestep hidden state, c_t is the t-th timestep LSTM memory cell state. $\sigma(\cdot)$ and $\tan h(\cdot)$ are sigmoid activation function and tanh activation, respectively. $\odot$ indicates the multiplication of matrix elements.

Ordinary LSTM can only capture semantic features in one direction. In this paper, we use bidirectional LSTM to collect semantic features in both directions, and the two representations are concatenated together as the output:

$$\begin{aligned} &\overrightarrow{h_t} = LSTM(e_t, \overrightarrow{h_{t-1}}), \overleftarrow{h_t} = LSTM(e_t, \overleftarrow{h_{t-1}}) \\ &h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \end{aligned} \tag{2}$$

3.3. Attention layer

Bi-LSTM still has the problem of semantic information loss for longer word order. Thus, we incorporate attention mechanisms that ignore distance limitation and obtain more long-distance semantic information. The output of Bi-LSTM is inputted to the attention layer. The current word is aligned with all the words in the sentence, and the result is normalized to obtain the weight of each word. Finally, a weighted summation of H is performed, which results in a representation that contains more semantic information. The calculation is as follows:

$$\begin{aligned} &a_{i,j} = \frac{\exp(f(h_i, h_j))}{\sum \exp(f(h_i, h_j))}, s_j = \sum_i a_{i,j} h_j \\ &F(h_i, h_j) = h_i^T \cdot h_j \end{aligned} \tag{3}$$

where the alignment function F is the dot product operation of two vectors and $a_{i,j}$ is the normalized weighting factor. In this way, the final output $S = \{s_1, s_2, ..., s_n\}$ is obtained.

3.4. CRF layer

As the last layer of the model, CRF layer performs sequence labeling on the output of attention layer $S = \{s_1, s_2, ..., s_n\}$ with S corresponding to the label $L = \{l_1, l_2, ..., l_n\}$. Initially, the state matrix $M \in R^{n \times m}$ and the transfer matrix $M \in R^{m \times m}$ are generated with S. M describes the mapping matrix between S and L, m is the number of sequence labels, and the data in the A matrix represent the probability of transferring from label y_i to label y_{i+1}. Then the score is calculated utilizing the matrices M and A, as follows:

$$\begin{aligned} &K(S, L) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} M_{i, y_i} \\ &p(L \mid S) = \frac{e^{K(S,L)}}{\sum_{L^* \in L_X} e^{K(S,L^*)}} \\ &\log(p(L \mid S)) = K(S, L) - \log(\sum_{L^* \in L_X} e^{K(S,L^*)}) \end{aligned} \tag{4}$$

L_X are all possible labels predicted by the model. The model is trained by maximizing the loss function, and the highest conditional probability can be obtained through training selection to label the input sequence.

4. Experiment

4.1. Dataset & Experimental settings

We crawl data from recruitment websites and construct a dataset. Then we use BMEO annotation method to segment the data at character-level. B represents the beginning of a word, M represents the middle part of a word, E represents the end of a word, and O indicates a word. After annotation, 16,840 pieces of job descriptions are obtained, 80% are selected as training data and 10% are selected as validate data, and the rest are for testing.

In this paper, BERT is applied to produce the word vector whose dimension is 768. A dropout layer is added between the word vector and Bi-LSTM, and the dropout ratio is set to 0.8. The hidden size of Bi-LSTM is set to 100. We use Adam optimizer with the learning rate set to 0.002 and the number of epoches set to 200.

4.2. Comparative methods

To confirm the experimental effectiveness of the BAC model, we compare our model with the following methods.
LSTM+CRF model: The model utilizes common LSTM and CRF layers, with the LSTM layer capturing semantic information and CRF layer annotating sequences.
Bi-LSTM+CRF model: The model is designed to use Bi-LSTM and CRF layers, with Bi-LSTM capturing semantic information and the CRF layer annotating sequences.
The comparison methods all use BERT to generate word vectors.

4.3. Evaluation index

This paper employs Precision(P), Recall(R), and F1-score as evaluation indicators. P indicates the samples' proportion with the correct prediction label among all samples. R indicates how much of the data in the sample is predicted correctly. F1 considers both precision and recall.

4.4. Experimental results

Table 1 shows the experimental results for BAC and baselines. BAC model achieves optimal results in all metrics. After the introduction of Bi-LSTM, Bi-LSTM+CRF improved significantly in performance over LSTM+CRF, proving that Bi-LSTM captures more semantic features through bi-direction. The performance is further boosted when attention mechanism is integrated, demonstrating that attention mechanism capture more semantic information.

Table 1. Experimental results.

Model	P	R	F1
LSTM+CRF	0.9103	0.8577	0.8832
Bi-LSTM+CRF	0.9100	0.8904	0.9001
BAC	**0.9277**	**0.8985**	**0.9129**

To validate the effectiveness of BERT, we use Word2vec[11] for contrast. The fundamental principle of Word2vec is to use a 3-layers neural network, by training large-scale data, extract semantic information between texts so that semantically similar words get similar vector expression. Initially, each word is randomly initialized to a low-dimensional vector trained to obtain the optimal word vector containing contextual semantic information.

As shown in Table 2, BERT is more superior than Word2vec, indicating that BERT can better capture the relationship between words, produce better word vectors and improve the model effect.

Table 2. Experimental results of different word embedding methods.

Model	Embedding	P	R	F1
BAC	Word2vec	**0.9353**	0.8857	0.9098
	BERT	0.9277	**0.8985**	**0.9129**

4.5. Parameter analysis

In this paper, we analyze the impact of the learning rate on the output results, and the result is shown in Fig. 2.

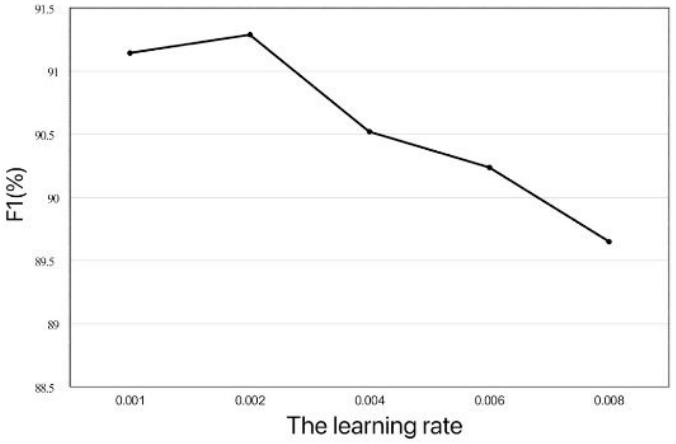

Fig. 2. Parameter analysis.

Five learning rates of 0.001, 0.002, 0.004, and 0.008 are trained on the model separately, and the model performance is measured with F1 value on the test set. Figure 2 reveals that the model performance initially increases and then decreases with increasing learning rate. When the learning rate is 0.002, the best model is learned. When the learning rate is low, the update step is small, the convergence rate is slow and the model takes a long time to reach a better result. If the learning rate is too high, the update step is large and the model converges fast. But it may skip local or global optimal solutions due to the large step size.

5. Conclusion

In this paper, we research on the Internet data entity extraction task and propose BAC model. Based on the CRF, this paper introduces bidirectional LSTM to extract semantic features. Meanwhile, to extract long-distance semantic feature, the attention mechanism is added. To validate the performance, we crawl the Internet recruitment data, and label the skill entities of the job description information. The experimental results show that the model achieves the optimal results. This paper compares the two word embedding methods of Word2vec and BERT. And we observe that the word vectors from BERT contain more semantic information. In future work, the BAC model will be applied to more complex areas like auditing. The documentation in the audit area is non-vertical and unstructured. The structured implementation of problem descriptions lays an important foundation for subsequent research on a matching system basis and generating audit opinions.

References

1. Prat F , Casacuberta F , María José Castro. Machine Translation with Grammar Association: Combining Neural Networks and Finite-State Models. *Proceedings of the Second Workshop on Natural Language Processing and Neural Networks* 2001.
2. Sabek I, Mokbel M F. Sya: Enabling Spatial Awareness inside Probabilistic Knowledge Base Construction. *ICDE* 2020.
3. Zhao Y, Zhang J, Xia X, et al. Evaluation of Google question-answering quality. *Library Hi Tech* 2019.
4. Lourenço H R, Martin O C, Stützle T. Iterated local search: Framework and applications. *In: Jones BS, Smith RZ, editors. Handbook of metaheuristics.* Springer Publishing Company Inc; 2019. p. 129-168.
5. Quimbaya A P, Múnera A S, Rivera R A G, et al. Named entity recognition over electronic health records through a combined dictionary-based approach. *Procedia Computer Science* 2016;**100**;55-6.
6. S Kim J H, Woodland P C. A rule-based named entity recognition system for speech input. *Sixth International Conference on Spoken Language Processing* 2000.
7. Li Y, Bontcheva K, Cunningham H. SVM based learning system for information extraction. *International Workshop on Deterministic and Statistical Methods in Machine Learning* 2004:319-20.
8. Szarvas G, Farkas R, Kocsor A. A multilingual named entity recognition system using boosting and c4. 5 decision tree learning algorithms. *International Conference on Discovery Science* 2006:267-11.

9. McCallum A, Li W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. Association for Computational Linguistics* 2003:188-2.
10. Kuru O, Can O A, Yuret D. Charner: Character-level named entity recognition. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* 2016:911-10.
11. Goldberg Y, Levy O. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv* 2014.
12. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv* 2014.
13. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv* 2015.
14. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation* 1997;**9(8)**:1735-45.
15. Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv* 2018.
16. Watanabe T, Tamura A, Ninomiya T, et al. Multi-Task Learning for Chemical Named Entity Recognition with Chemical Compound Paraphrasing. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* 2019:6245-5.
17. Xiao S, Ouyang Y, Rong W, et al. Similarity Based Auxiliary Classifier for Named Entity Recognition. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* 2019:1140-9.
18. Wang Z, Shang J, Liu L, et al. CrossWeigh: Training Named Entity Tagger from Imperfect Annotations. *arXiv preprint arXiv* 2019.
19. Liu T, Yao J G, Lin C Y. Towards improving neural named entity recognition with gazetteers. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* 2019:5301-6.