The Resume Corpus: A Large Dataset for Research in Information Extraction Systems

Yanyuan Su, Jian Zhang
School of Computer Science and Technology
Dongguan University of Technology
Dongguan, China
15602293616@163.com, zhangjian@dgut.edu.cn

Jianhao Lu
Shenzhen Intelligent Technology Co., Ltd.
Shenzhen, China
594830589@qq.com

***Abstract*—We publish a Chinese Resume Corpus for researches of information extraction. The corpus contains 178 thousand resume documents and over 33 million words. The resume documents with unstructured form contain many types of common information like name, gender, birthday, education, work experience, and so forth. We evaluate this corpus by using four types of mainstream neural network models for demonstrating the potential of our corpus to promote the development of information extraction research.**

Keywords-NLP; information extraction; resume corpus

I. INTRODUCTION

With the rapid development of Internet information technology, people are frequently exposed to text information in their daily lives, thus the text information has become the critical part of the Internet transmission data. There is a big challenge to extract useful information from massive amounts of data. To overcome the challenge, researchers have proposed several neural network-based algorithms for information extraction. However, if we want to train the powerful neural network models to implement information extraction automation technology, we need to collect a large number of dataset resources. Besides, the dataset resources of different application areas are hard to transfer from one area to another. Nowadays, as more and more social jobs are appearing, there is a question to how to quickly screen the resume documents from a large number of job seekers. If the automation technology can be applied on the resume screening work, it will improve the efficiency of work and save the cost. Hence, we publish the corpus which mainly focuses on the task of resume information extraction.

Natural language processing technology has been developing rapidly in recent years. Especially there are many deep learning algorithms proposed for information extraction in different application areas. Nowadays, Information extraction tasks for different fields are constantly appearing. Meanwhile, information extraction technology is also improving in order to pursue efficiency and accuracy. Z. Xie et al. [1] proposed a neural network model for medical-named entity recognition, which showed great potential in extracting medical-named entities from radiology reports. M. Q. Liang et al. [2] developed a named entity recognition tool for unstructured electronic prescriptions. Kun Yuan et al. [3] proposed a framework for effective information acquisition in social media. Rajbabu K. et al. [4] aimed to extract information from unstructured bidding documents of power plant industry. A binary classification method based on feature weight was proposed for information extraction in Rajbabu K. et al. However, there are few studies on resume information extraction. It is possible that they lack public resume information corpus, especially Chinese corpus. To fill the gap, we publish a corpus of Chinese resume[1] for information extraction research.

In order to evaluate the potential of this corpus, we perform a series of experiments to train and evaluate several neural network models on it. The models contain Convolutional Neural Network (CNN) [5], Bi-directional Long Short-Term Memory (BiLSTM) network [6], Bi-directional Gated Recurrent Unit (BiGRU) [8], and Bi-directional Encoder Representation from Transformers (BERT) [9]. We first take a partition of data from the corpus. We then manually label some tags on this part as labelled dataset. Finally, we use the cross-validation method to evaluate the potential of the corpus. We suppose that a corpus is useful if it can always get a stable result over different types of neural network models.

II. CORPUS CONSTRUCTION

Our resume information corpus is a large corpus that can meet data requirement of Chinese resume information extraction task. It has the following properties:

- Large number of data, including almost 178 thousand resume documents;
- Including common resume information: name, gender, date of birth, education, work experience, and so forth;
- As a Chinese corpus;
- Task-specific domain.

A. Data Collection

Our goal is to construct a large dataset of Chinese resume information to support the resume information extraction task. We crawl the resume documents from several Growth Enterprise Market Listed Companies that provide staff

[1] The dataset is available on: https://github.com/YanyuanSu/Resume-Corpus

978-1-7281-6092-4/19/$31.00 ©2019 IEEE
DOI 10.1109/CIS.2019.00087

profile information in their company announcement websites. The company announcements refer to the information of listed companies are released to the public through their designated platform, which contain open resume information of the staff. In these resume documents, there is personal basic information such as name, gender, place of origin and date of birth. Furthermore, there is personal experience information such as educational background and work history. In addition to the above information, there may be some resume documents containing other information such as awards, interests, personal strengths, etc.

B. *Corpus Creation*

We first extract the text of the resume information from the company announcement websites. Then, we perform a data cleaning operation to retain the information that can be used in information extraction tasks. After the series of processing, all resume information can be gathered into a corpus. An example of resume information in the corpus is shown in the Table VI.

C. *Corpus Content*

The attribute information of our Chinese resume information corpus is shown in the Table I. We find that the power of neural network models mainly depends on the size of training data resources. One of the important advantages of our corpus is its size, which is of great help to the research of resume information extraction based on neural network models. Another important feature is that the types of resume information in the corpus are very complete. In fact, there are series of common types of resume information in our corpus, which can effectively cope with downstream tasks in the research based on neural network models. Besides, our corpus is a Chinese corpus. It will contribute to the development of Chinese Natural Language Processing research.

D. *Annotation Dataset Generation*

We take out 1% of our corpus (randomly selected) to form an annotated dataset that can be used for evaluation of performance of the neural network models, which contains almost 1800 resume documents with a total of over 17000 sentences. We then split all resume documents into several sentences. There are different types of information in each resume sentence. In order to make a better distinction, we use six labels to represent different types of resume sentences. For example, some sentences contain personal basic information (name, gender, date of birth, nationality, ethnicity, etc.); some sentences contain education information (learning time, school, education background, degree, etc.); some sentences contain work experience information (work time, work location, department, position, etc.); some sentences do not contain any useful information. In this way, we annotate the sentences of resume as one or more labels from the six categories: personal basic information, past time, current time, study experience without time, work experience without time, and useless information, as shown in the Table II.

TABLE I. THE ATTRIBUTE INFORMATION OF RESUME CORPUS

Corpus Attributes	
#Resume Documents	177,214
#Sentences	1,741,045
#Words	33,700,541
#Txt Files	189

TABLE II. THE LABEL TYPES OF RESUME SENTENCES

The Types of Resume Sentences	Label
Personal Basic Information	basic
Past Time	ptime
Current Time	ctime
Study Experience Without Time	sexp
Work Experience Without Time	wexp
Useless Information	noinfo

To guarantee the annotation quality, we recruit four annotators including three undergraduate students and one master student. They all have a good knowledge in resume information classification. They are asked to independently annotate 100 examples first before labelling the whole dataset, with the aim to minimize ambiguity while strengthen the inter-annotator agreement. We define the standard of a sentence in terms of the label that receives the majority votes. We are interested in the necessary information of the resume. Thus, we first annotate the resume sentences with five common categories: personal basic information (basic), past time (ptime), current time (ctime), study experience without time (sexp), work experience without time (wexp).

In fact, there are some ambiguous categories during the annotation process. In order to remove these ambiguities, we discuss and decide on a uniform annotation standard before continuing annotation work. For example, a sentence for time information maybe contain classification ambiguity, the original sentence and the labeled sentence are shown in the Table III. In addition, there may be other information in the resume, such as awards information, project experience, positions held and so forth. However, this kind of information is not always necessary to appear in resume. In particular, there is no unified way to express this information for each resume. In order to demonstrate the generalization ability of research and reduce complexity of data, we classify this kind of information as the sixth category (noinfo) except the five categories mentioned above, as shown in the Table VII.

III. EVALUATION WITH CORPUS

This paper focuses on publishing a Chinese resume information corpus, and then demonstrates its potential to facilitate the training of information extraction models through the following experiments.

TABLE III. (A) THE ORIGINAL SENTENCE

ID	The Examples of Original Sentences
1	2016年毕业于北京理工大学网络工程。
2	毕业于山东省轻工业经济管理学校经济法律事务专业。
3	1992年7月至1994年8月就职于东港集团公司。
4	2014年11月至2016年4月，就职于和宁有限公司。

(B) THE LABELED SENTENCE

ID	The Examples of Labeled Sentences
1	2016年毕业于北京理工大学网络工程[ptime]
2	毕业于山东省轻工业经济管理学校经济法律事务专业[sexp]
3	1992年7月至1994年8月就职于东港集团公司[ptime]
4	2014年11月至2016年4月[ptime] 就职于和宁有限公司[wexp]

TABLE IV. THE PARAMETERS OF THE EXPERIMENTAL MODEL

Model	Parameters
CNN	filters=100, kernel_size=4, padding='same', use Global Average Pooling.
BiLSTM	units=400，dropout=0.5.
BiGRU	units=400，dropout=0.5.
BERT	max_seq_length=50, batch_size=200, learning_rate=0.001, dropout=0.5.

TABLE V. THE RESULT OF THE EXPERIMENTS

Eigenvector	Model	F1	Accuracy
BERT-Base, Chinese	BERT+CNN	0.926	0.963
	BERT+BiLSTM	**0.958**	**0.987**
	BERT+BiGRU	0.954	0.986
	BERT	0.929	0.974
Tencent AI Lab Embedding Corpus for Chinese Words and Phrases	CNN	0.583	0.877
	BiLSTM	0.823	0.929
	BiGRU	0.815	0.926

A. Experimental Settings

We run all baselines on a classification task and predict the sentence label of each resume document with several types of neural network models. In order to predict the labels of all resume sentences, we use the feature vector of pre-training model to convert the sentences into the feature vectors. We then use the supervised learning algorithms to train the classification models, which can predict the labels of the resume sentences.

Comparative models (1) **CNN**: We employ a CNN including a convolutional layer, a pooling layer, a fully connected layer [5]. (2) **BiLSTM**: BiLSTM is a neural network model that consists of forward LSTM and backward LSTM. LSTM can better capture long-distance dependencies [7]. Because of its design features, it is ideal for modeling the time series data, for example the text data [6]. Considering that when using LSTM, we can't encode the information from back to front, so we decide to use BiLSTM with bidirectional function. (3) **BiGRU**: GRU can be seen as a variant of LSTM, which replaces the forget gates and input gates in LSTM with updated gates [8]. (4) **BERT**: BERT is a method of pre-training language representations published by Google [9]. It has achieved state-of-the-art scores in many natural language processing tasks. We can use its pre-training model directly for classification tasks. The parameters of the experimental model are shown in the Table IV, which is default parameter if not mentioned.

Eigenvector We employ two methods of eigenvector: (1) **BERT-Base, Chinese**: It is the Google's official pre-training Chinese model, which contains Chinese Simplified and Traditional, 12-layer, 768-hidden, 12-heads, 110M parameters, and we use it to generate the sentence embedding. (2) **Tencent AI Lab Embedding Corpus for Chinese Words and Phrases [10]**: This corpus provides 200-dimension vector representations, a.k.a. embeddings, for over 8 million Chinese words and phrases, which are pre-trained on large-scale high-quality data. Before the process of Tencent's word embedding, we use Jieba[2], which is a Chinese word segmentation module to divide resume sentences into words.

Evaluation metrics We divide the annotation dataset into three parts: (1) the training dataset (60% of the annotation dataset, contains over 1000 resumes), (2) the verification dataset (20% of the annotation dataset, contains over 300 resumes), (3) the test dataset (20% of the annotation dataset, contains over 300 resumes). In order to achieve more convincing results by using the dataset, we perform the experiments using 5-fold cross-validation for all comparative models. We adopt **F1-score** and **accuracy** as the evaluation metrics. Accuracy is simple and intuitive, but it is not a truly correct indicator in many cases. Thus, we add F1-score to the evaluation metric [11]. In statistical analysis of binary classification, the F1-score is a measure of a test's accuracy, which considers both the precision and the recall of the test to compute the score.

B. Empirical Result

The results of the experiments are shown in the Table V. The best performing system is indicated in bold.

We find that the models based on sentence embedding is superior to the models based on word embedding on all evaluation metrics. It is probably that sentence embedding can catch more sematic information to help classification process. In addition, we find that BiLSTM achieves the best results, and BiGRU as its variant has achieved similar results. This is most likely the effect of their abilities to capture long-distance dependencies [12]. Besides, there is a large difference in the F1-score and accuracy of CNN based on word embedding. It is probably due to the limited number of neural network layers of standard CNN, which can be solved

[2] J Sun, "'Jieba'Chinese word segmentation tool." 2012 https://github. com/fxsjy/jieba

by adding the convolutional layer of CNN. Finally, the excellent F1-score and accuracy rates obtained by the BiLSTM and BiGRU models demonstrate that the potential of our corpus to downstream tasks.

IV. CONCLUSIONS

We published the first open Chinese resume information corpus and described the construction of the corpus and its properties. The purpose of the Chinese resume information corpus is to provide an open resource that can be applied to the Chinese information extraction research area in the era of rapid development of information extraction technology. In addition, we performed a series of experiments for utility of the corpus. The results of four neural network models trained on the corpus has demonstrated the potential for our corpus. We will further design a serials of information extraction models based on the corpus in our future work.

ACKNOWLEDGMENT

This work was supported by the Natural Science Foundation of China (Grant No.61631016).

REFERENCES

[1] Z. Xie, Y. Yang, M. Wang, M. Li, H. Huang, D. Zheng, ... and T. Ling, “Introducing information extraction to radiology information systems to improve the efficiency on reading reports”, *Methods of Information in Medicine 58*, pp. 094-106, 2019.

[2] M. Q. Liang, V. Gidla, A. Verma, D. Weir, R. Tamblyn, D. Buckeridge and A. Motulsky, “Development of a Method for Extracting Structured Dose Information from Free-Text Electronic Prescriptions”, *Studies in health technology and informatics 264*, pp.1568-1569, 2019.

[3] Kun Yuan, Guannan Liu, and Junjie Wu, "Whose posts to read: Finding social sensors for effective information acquisition", *Information Processing & Management* 56, pp. 1204-1219, 2019.

[4] K. Rajbabu, H. Srinivas and S. Sudha, “Industrial information extraction through multi-phase classification using ontology for unstructured documents”, *Computers in Industry 100*, pp.137-147, 2018.

[5] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, “Gradient-based learning applied to document recognition”, *Proceedings of the IEEE 86*, pp.2278-2324, 1998.

[6] S. Hochreiter and J. Schmidhuber, “Long short-term memory”, *Neural computation 9*, pp.1735-1780, 1997.

[7] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks”, *IEEE Transactions on Signal Processing 45*, pp.2673-2681, 1997.

[8] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation” *arXiv preprint arXiv:1406.1078*, 2014.

[9] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding”, *arXiv preprint arXiv:1810.04805*, 2018.

[10] Yan Song, Shuming Shi, Jing Li, and Haisong Zhang, “Directional skip-gram: Explicitly distinguishing left and right context for word embeddings”, In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 175-180, 2018.

[11] Xiao Li, Ye-Yi Wang, and Alex Acero, "Learning query intent from regularized click graphs", Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2008.

[12] Alex Graves, Santiago Fernández, and Jürgen Schmidhuber, "Bidirectional LSTM networks for improved phoneme classification and recognition", *International Conference on Artificial Neural Networks*, Springer, Berlin, Heidelberg, 2005.

TABLE VI. EXAMPLE OF RESUME INFORMATION IN THE CORPUS

1 王泉庚,男,汉族,1972年10月出生,中国籍,无境外永久居留权,毕业于中欧国际工商学院工商管理硕士,巴黎国际时装艺术学院艺术管理专业,硕士学位。1995年11月至2013年11月,就职于上海美特斯邦威服饰股份有限公司,历任副总经理、董事兼副总裁;2013年12月至2014年2月,离职调整;2014年3月至今,就职于时朗企业发展(上海)有限公司,任执行董事兼总经理;2014年8月至2015年5月,就职于好孩子(中国)商贸有限公司,任执行董事兼总经理;2015年4月至今,任南京我乐家居股份有限公司独立董事;2016年4月至今,任迅驰时尚(上海)科技股份有限公司董事。
2 张诗琪,女,汉族,1981年2月出生,中国籍,无境外永久居留权,毕业于上海大学国际英语与国际贸易系专业,本科学历。2001年8月至2004年10月,就职于智尚(上海)品牌管理有限公司,任客户部经理;2004年11月至2008年7月,就职于上海迅驰广告有限公司,任运营总监;2008年8月至2016年3月,就职于上海迅驰时尚品牌管理有限公司,历任运营管理部总监、执行董事兼总经理、法定代表人;2016年4月至今,就职于迅驰时尚(上海)科技股份有限公司,现任董事、副总经理。

TABLE VII. EXAMPLE OF ANNOTATED SENTENCES

Unlabeled	方涛,男,汉族,1979年1月出生,中国籍,无境外永久居留权,毕业于上海交通大学安泰经济与管理学院工商管理专业,硕士学位。2003年7月至2008年7月,就职于上海迅驰广告有限公司,任总经理。
Labeled	方涛[basic] 男[basic] 汉族[basic] 1979年1月出生[basic] 中国籍[basic] 硕士学位[basic] 毕业于上海交通大学安泰经济与管理学院工商管理专业[sexp] 2003年7月至2008年7月[ptime] 就职于上海迅驰广告有限公司[wexp] 任总经理[wexp] 无境外永久居留权[noinfo]