Fast and Accurate Resume Parsing Method Based on Multi-task Learning

Yuhai Liang
Dept. of S3- HR and Management Line
Tencent
Shenzhen, China
johanliang@tencent.com

Zhicheng Lin
Dept. of S3- HR and Management Line
Tencent
Shenzhen, China
essienlin@tencent.com

Xin Shi
Dept. of S3- HR and Management Line
Tencent
Shenzhen, China
hermanshi@tencent.com

Haigang Ma
Dept. of S3- HR and Management Line
Tencent
Shenzhen, China
haigangma@tencent.com

***Abstract*—With the growing demand for online recruitment, an efficient and highly accurate resume parsing method is needed. Existing resume parsing methods are basically two-stage approaches: resume segmentation and entity recognition. Some researchers apply deep learning to these two stages to improve the accuracy but at the cost of lower efficiency. In this article, we propose a multi-task deep learning model that completes resume segmentation and entity recognition at the same time, thus reducing the resume parsing time by nearly half. Moreover, our method can better parse the dual-list format resumes, which is usually ignored by other existing methods. We also use some post-correction rules to further improve the accuracy of resume parsing. Finally, compared with the baselines, we achieve state-of-the-art on both Chinese and English resumes.**

Keywords—resume parsing, resume segmentation, entity recognition, multi-task learning, dual-list format resume.

I. Introduction

With the rapid growth of the Internet and digital technology, the volume of resumes and candidate information is growing exponentially. Traditional methods of manually extracting information from resumes are no longer applicable [1]. Resume parsing technology aims to extract and organize the information in a resume document automatically, making the recruitment process more efficient. Besides, with results obtained from resume parsing, the recruitment system can develop other helpful functions, such as a resume search engine [2], resume ranking [3], and position recommendation [4]. Therefore, resume parsing is a fundamental and critical function in a recruitment system. So, requirements for resume parsing technology are getting higher, not only high efficiency but also high accuracy are required.

Resume content is usually semi-structured and consists of multiple different topic segments, such as “Basic Information”, “Expectation Information”, “Education Information”, “Work Information”, “Project Information”, “Summary”, etc., but their order and layout structure are changeable [5]. In addition to the common list format, there are also dual-list format, tabular format, or unordered blob-based format [6]. Especially when the resume text is converted from other file formats, such as pdf, doc, docx, etc., unexpected formats may appear [7]. Current approaches normally assume a standard format which they can parse, or assume that the parsed information is available to use for their task [6].

Usually, researchers divide resume parsing into three steps. The first step is to segment the resume content according to different topics. The second step is to extract target fields from these topic segments. Finally, form the structured data based on the topic segments and the fields. The first two steps are the key to resume parsing.

The topic segment in a resume generally starts with a heading. For the common list format resumes, all of the text information available between the start and the end of the heading is accepted as a segment [8]. Therefore, researchers usually consider the resume segmentation as a task of text matching or text classification for each line of the resume to identify headings. However, if part of the resume content is in the dual-list format, it will cause a segmentation error because lines in the dual-list format usually have two different topics and cannot be accepted as a segment. Many methods used to parse the list format resume are not suitable for the dual-list format resume, leading to low accuracy.

After resume segmentation, according to the topic of each segment, the corresponding rule-based regular expressions and keyword-matching techniques are used to extract the target fields [7]. At present, Named Entity Recognition (NER) has seen significant advancements in information extraction, from traditional models like Hidden Markov Models [9] and Conditional Random Fields (CRF) [10] to recent deep learning models such as BERT-BiLSTM-CRF [11]. So, extracting the target fields in the topic segments is regarded as an NER task by researchers.

Finally, entities obtained from NER can be mapped to target fields according to their respective topic segments. Then the resume can be well structured. However, in most current resume parsing solutions, resume segmentation, and entity recognition are seen as two stages, executed serially. This affects the speed of resume parsing, especially when these two stages use computationally intensive large deep learning models. Therefore, we parallel these two tasks in a way of training a multi-task deep learning model, which can segment the resume text and recognize entities at the same time.

First, we concatenate each line with its upper and lower lines using token [CLS] and token [SEP] in the form of “[CLS] current line [SEP] upper line [SEP] lower line [SEP]” as the input of the BERT model [12], which is a large pre-training model trained on a massive dataset of unlabeled text and has strong transfer learning ability. The output of the BERT model is a sequence of word embeddings. Then the embeddings of the [CLS] token and the first [SEP] token will be used to classify the topic category of the current line. If the line is the list format content, the topic categories of the [CLS] token embedding and the first [SEP] token embedding will be the same. If it is the dual-list format content, the categories of the

[CLS] token embedding and the first [SEP] token embedding will be different. With the predicted topic categories of all lines, consecutive lines with the same topic category can be regarded as a topic segment. For the dual-list format content, we only need to do this by list to get all the topic segments. At the same time, all the word embeddings will be exported to the BiLSTM-CRF [13] network to predict the sequence of entity tags. Finally, the entities can be mapped to the target fields according to their respective topic segments.

In summary, in this paper, we have made the following three contributions:

1) The multi-task learning model we proposed can complete topic classification and entity recognition of the resume line at the same time, which greatly improves the efficiency of resume parsing.

2) We provide a new and feasible solution for parsing the dual-list format resume, which is ignored by other existing methods.

3) Compared with baselines, our method achieves state-of-the-art on both Chinese and English resumes.

The rest of the paper is organized as follows. Section 2 introduces the related work of resume parsing. Section 3 illustrates the architecture of our model in detail. Section 4 describes the experiments and analyses the results. Finally, we present a brief conclusion in section 5.

II. Related Work

Various methods have been proposed to extract structured information from unstructured resumes, including rule-based methods, machine learning-based methods, and deep learning-based methods.

Varsha Tiwari et al. [7] store common headings as a data dictionary which is used to match the headings in a given resume to find segments. Then for each segment, there is a group of entity dictionaries and regular expressions to match the target fields. Similarly, Al-Amoudi et al. [14] parse the resume file into XML format and then use rule-based regular expressions and keyword-matching techniques to extract metadata. This kind of rule-based method requires a lot of manual labor and cannot handle diverse resume formats.

Machine learning-based methods can learn the structure and features of resumes automatically without writing rules manually [15]. Gunaseelan, B. et al. [16] use Extreme Gradient Boosting (XGB) [17], to classify the headings of a resume. All lines are labeled as two target classes: ‘1’ for a heading and ‘0’ for not a heading. Then fuzzy string matching is utilized to extract the skillsets. Kun Yu et al. [18] apply HMM for the resume segmentation and apply SVM [19] for information extraction. PROSPECT [20] is a system for screening candidates for recruitment, which uses CRFs for resume segmentation and then the segmented textual sections are parsed by CRF-based extractors to derive named entities. The accuracies of the methods are not high enough because of the weak representation ability of machine learning.

Compared to machine learning methods, deep learning has more powerful representation learning capabilities, which can better represent features without feature engineering. Shicheng Zu et al. [21] use Attention BLSTM as the line label classifier to segment text, achieving an $F1$-score of 0.86. Then BiLSTM-CNNs-CRF is employed for entity recognition, achieving 0.839 $F1$-score. When BERT [12] was proposed, Nirmiti Bhoir et al. [22] utilized the techniques of fuzzy string matching and n-grams to match headings in resumes for resume segmentation and fine-tuned the BERT model on a labeled dataset of resumes, finally achieving an accuracy of 90-92% for different entities. Wang et al. [23] used BERT to encode resume texts into word vectors and then exported them to BiLSTM-CRF to produce the entity tag sequence, which can reach the F-1 value of 94. 82% of entity recognition. Although the use of deep learning models for resume segmentation and entity recognition can further improve the parsing effect, the speed decreases, especially when the large pre-training model, like BERT, is used for both tasks.

As for the dual-list format resume content, Bhatia et al. [5] apply heuristic rules based on the spacing of text across the page to tell the difference between the list format content and the dual-list format content. Larger spaces between words on the same horizontal line would indicate inter-column gaps whereas smaller spaces would indicate normal inter-word gaps. However, large spaces also often appear in the line of list format contents, and spatial structure is often changed when retrieving resume text from files [7]. Our approach identifies the dual-list format content based on the semantics rather than the spatial structure.

III. Method

We design a multi-task learning model that can perform topic classification and entity recognition for the line of the resume. Some rules based on prior knowledge are set to post-correct the results of the topic classification. Finally, the resume is segmented according to these topic categories, and entities are mapped to target fields to structure the resume.

A. Topic Classification

Fig. 1. shows the overall structure of our model. With the help of contextual information, it classifies the topic category of each line of the resume. The input in the form of “[CLS] current line [SEP] upper line [SEP] lower line [SEP]” will be encoded by BERT, outputting a feature vector for every single token, which is also called word embedding. The [CLS] token embedding is usually applied for the task of sentence classification [12], therefore it is used to classify the topic category of the left part of the resume line. The first [SEP] token embedding is used to classify the right part of the resume line. These two predicted topic categories from the trained model are consistent for the list format content, while different for the dual-list format content.

The preset topic categories that the model learns are as follows: “Basics_heading”, “Basics”, “Expectation_heading”,

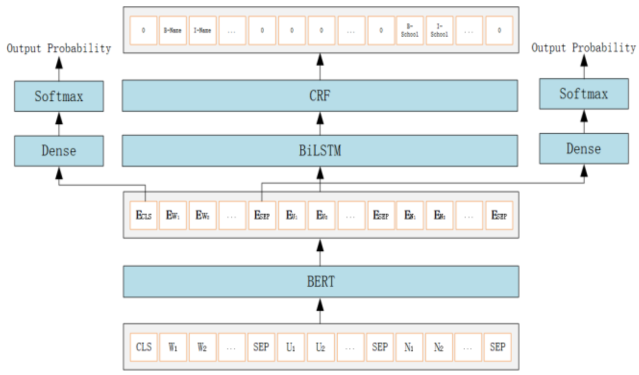

Fig. 1. The Structure of the Model

“Expectation”, “Education_heading”, “Education”, “Work_heading”, “Work”, “Project_heading”, “Project”, “Summary_heading”, “Summary”. The category with the suffix “_heading” indicates the heading of the topic segment. “Summary” is additional introductory information about the candidate.

B. *Entity Recognition*

As shown in Fig. 1, while categorizing the topic of the current line, all word embeddings will be input to the downstream BiLSTM-CRF network for the prediction of the sequence of entity tags. Note that only the entities before the position of the first [SEP] token belong to the current line and should be retained.

The preset entities that the model learns are as follows: Name, Address, Email, Phone, Date, School, Major, Degree, Company, Position, Project Name, Skill, and Language.

C. *Post-Correction*

Errors in topic categories will affect subsequent resume segmentation, resulting in poor final parsing. With the predicted results of the multi-task learning model, we propose some post-correction rules to further improve the accuracy of the topic classification results. The rules are as follows:

- If the entities in the current line are School and Major, the topic category should be “Education”.
- If the entities in the current line are School and Degree, the topic category should be “Education”.
- If the entities in the current line are School and Date, the topic category should be “Education”.
- If the entities in the current line are Major and Degree, the topic category should be “Education”.
- If the entities in the current line are Company and Position, the topic category should be “Work”.
- If the entities in the current line are Company and Date, the topic category should be “Work”.
- If the entities in the current line are Project Name and Date, the topic category should be “Project”.
- If the previous heading is “Education_heading”, the current topic category should be “Education”.
- If the previous heading is “Work_heading”, the current category should be “Work”.
- If the previous heading is “Project_heading”, the current category should be “Project”.
- If the previous heading is “Summary_heading”, the current category should be “Project”.
- If the topic categories of the upper line are the same as those of the lower line, the current topic category should be the same as them.

The predicted [CLS] topic categories and [SEP] topic categories of lines will be corrected by the above rules.

D. *Structure*

Fig. 2 shows the process of the resume parsing. On both sides are predicted [CLS] topic categories and [SEP] topic categories, while in the small box are the recognized entities. Post-correction rules can recorrect some topic errors. Then the

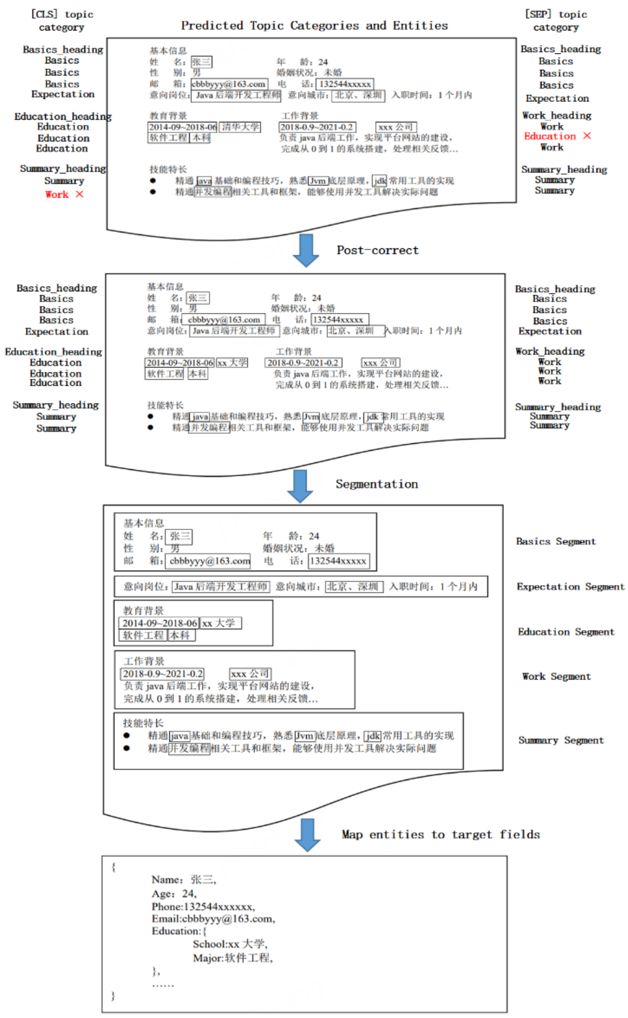

Fig. 2. The Process of Resume Parsing

consecutive lines with the same topic category can be treated as a segment. For the line with two different topic categories, just split the line into two lists with the largest space within it and get two segments. After the resume segmentation, the recognized entities in segments can be mapped to the target fields and form structured data in the form of json.

IV. Experiment

A. *Data Preparation*

We have prepared 5000 Chinese resumes and 5000 English resumes and converted them to text with some software tools such as pdf-plumber, docx2txt, OCR, etc. As Table 1 shown, we manually annotate each line with [CLS] topic category, [SEP] topic category, and BIO sequence. The upper part of Table 1 is annotations for a list format resume, with the same annotation for the [CLS] topic and the [SEP] topic. The lower half is annotations for a dual-list format resume, and the annotations for the [CLS] topic and the [SEP] topic are different. The BIO annotation is a common way of annotating the NER task [24]. “B” is used to mark the first character of the entity, “I” is used to mark other characters of the entity, and “O” is used to mark non-entity characters.

Finally, the annotated resumes are randomly sampled into the training set, development set, and test set according to the ratio 8:1:1.

To present the dataset more clearly, Fig. 3 displays the distribution of different topic categories and entities on the

TABLE I. A PART OF ANNOTATED DATASET

[CLS] Topic	Resume & BIO Sequence		[SEP] Topic
Basics_heading	个人信息		Basics_heading
	O O O O		
Basics	张三 邮箱：johan@xxx.com		Basics
	B-Name I-Name … B-Email …		
Education_heading	教育经历		Education_heading
	O O O O		
Education	北京邮电大学 信息工程		Education
	B-School I-School … B-Major I-M...		
…	…		…
Basics	李四	工作经历	Work_heading
	B-Name I-Name O…	O O O O	
Basics	联系方式：13…	腾讯 软件开发	Work
	O O O O B-Phone …	B-Company …	
Basics	居住地：北京…	负责开发相关…	Work
	O O O O B-Address …	O O O O O…	

Chinese and English datasets. In the distribution of topic categories, the proportion of each heading category is relatively small, basically only accounting for about 1%. This is because generally different topic headings can appear at most once in a resume. Education and Work account for the largest proportion. Moreover, in the distribution of entities, the distribution of various entities is relatively balanced. Skill, Date, and Company entities account for the most. Finally, no matter the topic categories or entities, the distributions on Chinese and English datasets are similar.

B. *Training Details*

We employ the BERT-Base-Case model[1] for English resume data and the BERT-Base-Chinese model[2] for Chinese resume data. Our models are trained with the batch size of 8, the learning rate of 2e-4, the gradient clip value of 5, the dropout rate of 0.5, and the AdamW optimizer.

C. *Baselines*

In order to make a comprehensive comparison, a machine learning-based method and two deep learning-based methods are selected as baselines:

PROSPECT It is a recruitment system, which uses CRF [10], a machine learning algorithm, for resume segmentation and entity recognition. Since the system only supports English resumes and only the final parsing results can be seen, we only count the final target fields of the English resume parsing results.

Attention BiLSTM+BiLSTM-CNNs-CRF It is a state-of-the-art model for resume parsing [21], using Attention BiLSTM to classify the topic of every line, followed by BiLSTM-CNNs-CRF as the entity recognition network. The dual-list format content of the resume are ignored here, so during the training and testing, only the labeled [CLS] topic categories are used.

BERT+BERT-BiLSTM-CRF The only difference between this baseline and our method is that its models are separated and ours is a multi-task learning model. The input will be the same as ours and the [CLS] token embedding and the [SEP] token embedding of BERT are used for the task of

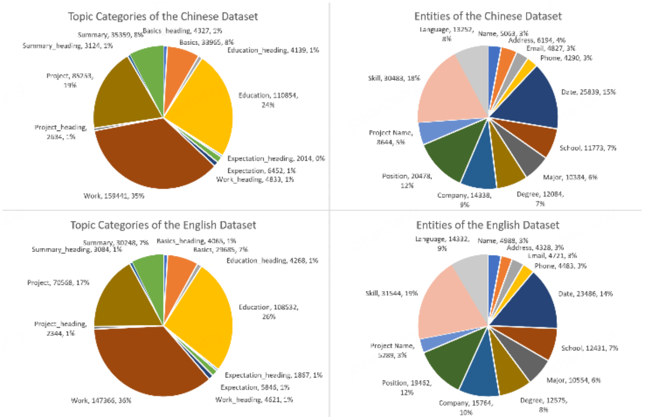

Fig. 3. The Distribution of Different Topics and Entities on the Chinese and English Datasets

topic classification to handle the dual-list format content. Then BERT-BiLSTM-CRF is used for entity recognition. For a fair comparison, the post-correction rules mentioned above are also used. For the topic classification task, both the predicted categories of [CLS] and [SEP] will be used for evaluation.

All the BERT models here are BERT-Base-Case and the BERT-Base-Chinese model.

D. *Evaluation Metrics*

We use three metrics: *Precision*, *Recall*, and F_1 *score* to evaluate the topic classification performance, and only F_1 *score* is used for the task of entity recognition for saving space.

E. *Experimental Results*

The topic classification experimental results are listed in Table 2. "A+B" represents the baseline **A**ttention BiLSTM+**B**iLSTM-CNNs-CRF and "B+B" represents **B**ERT+**B**ERT-BiLSTM-CRF. Our model performs better in most topic categories on both Chinese and English resumes, achieving 0.88 of the average $F1$-score on the Chinese resumes and 0.87 of the average $F1$-score on the English resumes, higher than those of "A+B" and "B+B".

Moreover, the performance of headings is better than that of non-headings, indicating headings are easier to classify. "B+B" performs better than "A+B", indicating that to some extent, the transfer learning ability of the BERT model is helpful for the topic classification of resumes. Our model performs better than "B+B", indicating identifying the topic could benefit from identifying entities.

Table 3 shows the $F1$-score of the target fields and the speeds of various resume parsing methods. Our model achieves 0.933 of the $F1$-score of the Chinese resumes and 0.929 of the $F1$-score of the English resumes, higher than other models. Likewise, our model performs better than "B+B", suggesting that multi-task learning is helpful for the entity recognition task.

Moreover, concerning the resume parsing speed, we set the speed of our method to be the baseline and compared it with the various methods. Although the $F1$-score of PROSPECT based on the machine learning model has a large gap with other deep learning methods, the speed is the fastest.

[1] https://github.com/google-research/bert/ BERT-Base, Cased
[2] https://github.com/google-research/bert/ BERT-Base, Chinese

TABLE II. EXPERIMENTAL RESULTS OF TOPIC CLASSIFICATION

Topic Category	Model	Chinese Resume			English Resume		
		P	R	F_1	P	R	F_1
Basics_ heading	A+B	0.92	0.86	0.89	0.93	0.88	0.90
	B+B	0.94	0.87	0.91	0.94	0.88	0.90
	Ours	**0.96**	**0.89**	**0.93**	**0.95**	**0.90**	**0.92**
Basics	A+B	0.91	0.83	0.87	0.90	0.82	0.86
	B+B	0.92	0.84	0.89	0.93	0.84	0.88
	Ours	**0.93**	**0.86**	**0.90**	**0.94**	**0.86**	**0.90**
Expectation_heading	A+B	0.89	0.83	0.86	0.88	0.84	0.86
	B+B	0.91	0.86	0.88	**0.91**	**0.86**	**0.88**
	Ours	**0.91**	**0.87**	**0.89**	0.90	0.87	0.88
Expectation	A+B	0.91	0.84	0.87	0.90	0.83	0.86
	B+B	0.93	0.86	0.90	0.91	0.83	0.87
	Ours	**0.94**	**0.88**	**0.91**	**0.92**	**0.86**	**0.89**
Education_heading	A+B	0.93	0.87	0.90	0.93	0.88	0.90
	B+B	**0.95**	**0.89**	**0.92**	0.95	0.90	0.92
	Ours	0.94	0.89	0.91	**0.95**	**0.90**	**0.93**
Education	A+B	0.78	0.75	0.76	0.78	0.74	0.76
	B+B	0.84	0.80	0.82	0.83	0.79	0.81
	Ours	**0.88**	**0.83**	**0.86**	**0.88**	**0.84**	**0.86**
Work_ heading	A+B	0.89	0.85	0.87	0.90	0.86	0.88
	B+B	0.92	0.88	0.90	0.91	0.88	0.89
	Ours	**0.93**	**0.89**	**0.91**	**0.93**	**0.89**	**0.91**
Work	A+B	0.76	0.70	0.73	0.75	0.70	0.72
	B+B	0.86	0.85	0.85	0.87	0.84	0.86
	Ours	**0.89**	**0.87**	**0.88**	**0.88**	**0.84**	**0.86**
Project_heading	A+B	0.90	0.85	0.87	0.89	0.83	0.86
	B+B	0.92	0.88	0.90	0.93	0.89	0.91
	Ours	**0.94**	**0.88**	**0.91**	**0.95**	**0.88**	**0.91**
Project	A+B	0.78	0.71	0.74	0.77	0.70	0.73
	B+B	**0.86**	**0.78**	**0.82**	**0.85**	**0.77**	**0.81**
	Ours	0.86	0.77	0.81	0.84	0.76	0.80
Summary_heading	A+B	0.89	0.83	0.86	0.90	0.83	0.85
	B+B	0.91	0.85	0.88	0.92	0.86	0.89
	Ours	**0.93**	**0.86**	**0.89**	**0.93**	**0.86**	**0.89**
Summary	A+B	0.74	0.69	0.71	0.74	0.70	0.72
	B+B	0.85	0.77	0.81	0.86	0.78	0.82
	Ours	**0.88**	**0.76**	**0.82**	**0.88**	**0.78**	**0.83**
Avg.	A+B	0.86	0.80	0.83	0.86	0.80	0.83
	B+B	0.90	0.84	0.87	0.90	0.84	0.87
	Ours	**0.92**	**0.85**	**0.88**	**0.91**	**0.85**	**0.87**

"B+B" is slower than "A+B" because of the use of the large pre-train model BERT. Our method is almost twice as fast as "B+B" because the topic classification and the entity recognition of "B+B" are serial, while ours are simultaneous.

F. Ablation Analysis

To further confirm the contribution of each component of our model, we conduct ablation experiments. Table 4 shows the results.

Without post-correction rules, the average $F1$-score of the target field drops by 0.2 on Chinese resumes and 0.24 on English resumes, which means the post-correction is a helpful step.

Without [SEP] topic classification, parsing the dual-list

TABLE III. RESULTS OF TARGET FIELDS AND PARSING SPEED

Field	PROSPECT	A+B		B+B		Ours	
	en	zh	en	zh	en	zh	en
Name	/	0.893	0.892	0.940	0.951	0.944	0.953
Address	/	0.824	0.768	0.897	0.882	0.913	0.908
Email	0.864	0.925	0.918	0.983	0.978	0.994	0.988
Phone	0.804	0.930	0.922	0.980	0.982	0.980	0.982
Date	0.746	0.818	0.781	0.912	0.874	0.939	0.883
School	0.708	0.847	0.863	0.921	0.934	0.947	0.949
Major	/	0.804	0.813	0.890	0.901	0.916	0.922
Degree	0.834	0.885	0.858	0.955	0.935	0.958	0.942
Company	0.683	0.796	0.835	0.886	0.904	0.897	0.909
Position	0.717	0.783	0.800	0.875	0.887	0.891	0.902
Project Name	/	0.766	0.798	0.845	0.864	0.866	0.874
Skill	0.801	0.864	0.857	0.937	0.929	0.944	0.935
Language	/	0.878	0.860	0.939	0.922	0.941	0.930
Avg.	0.770	0.846	0.843	0.920	0.919	**0.933**	**0.929**
Speed	0.21×	1.35×		1.9×		1×	

format content results in more errors, decreasing the average $F1$-score by more than 0.3. And we can expect that the more dual-list format contents in the test set, the worse the performance will be. Therefore, the identification of dual-list format contents is very important.

Without the input of the upper and lower lines, the performance drops even more because it will reduce the accuracy of topic classification, which has an impact on the parsing of both list format and dual-list format contents.

After removing the post-correction, [SEP] topic classification, and the input of upper and lower lines, the $F1$-score of the Chinese resumes drops to 0.876 and that of the English resumes drops to 0.873.

In a word, the ablation study illustrates that the post-correction, the identification of the dual-list format content, and the contextual information are beneficial to resume parsing.

V. CONCLUSION

Resume parsing is the most fundamental function in the online recruitment system, which requires not only high accuracy but also fast parsing speed. We propose a multi-task learning model that can simultaneously perform topic classification and entity recognition for the line of the resume, thereby effectively segmenting resumes and mapping the identified entities to the target fields, ultimately obtaining structured resumes.

In addition, our method supports the parsing of the dual-list format content of the resume, which is rarely mentioned in other existing resume parsing methods. Moreover, we set

TABLE IV. RESULTS OF ABLATION ANALYSIS

Method	Chinese	English
Ours	**0.933**	**0.929**
- post-correction	0.913	0.905
- [SEP] topic classification	0.901	0.897
- upper and lower lines	0.894	0.891
- *above three*	0.876	0.873

some ingenious post-correction rules to further improve the parsing performance. Finally, in our experiments, our method achieved SOTA performance on both Chinese and English resumes, reaching the $F1$-score of 0.933 on the Chinese resume dataset and 0.929 on the English resume dataset.

In the future, we will continue to explore how to better parse other formats of resumes, including table resumes, unordered blob-based formats, and so on. At the same time, we will constantly improve the accuracy and speed of resume parsing.

REFERENCES

[1] Liu, Yijing. "IntelliExtract: An End-to-End Framework for Chinese Resume Information Extraction from Document Images." Advances in Engineering Technology Research 6.1 (2023): 570-570.

[2] Du, Jianping, and Dongping Ma. "Intelligent Resume Retrieval Based on Lucence." J. Softw. 17.1 (2022): 29-35.

[3] Guo, Shiqiang, Folami Alamudun, and Tracy Hammond. "RésuMatcher: A personalized résumé-job matching system." Expert Systems with Applications 60 (2016): 169-182.

[4] Chavare, Dipti Suhas, and Archana Bhaskar Patil. "Resume Parsing using Natural Language Processing." Grenze International Journal of Engineering & Technology (GIJET) 9.1 (2023).

[5] Z. Chuang, W. Ming, L. C. Guang, X. Bo and L. Zhi-qing, "Resume Parser: Semi-structured Chinese Document Analysis," 2009 WRI World Congress on Computer Science and Information Engineering, Los Angeles, CA, USA, 2009, pp. 12-16, doi: 10.1109/CSIE.2009.562.

[6] Bhatia, Vedant, et al. "End-to-end resume parsing and finding candidates for a job description using bert." arXiv preprint arXiv:1910.03089 (2019).

[7] Tiwari, Varsha, and Sapna Jain Choudhary. "Intelligent Hiring with Resume Parser and Ranking Using Machine Learning and Natural Language Processing."

[8] Barducci, Alessandro, et al. "An end-to-end framework for information extraction from italian resumes." Expert Systems with Applications 210 (2022): 118487.

[9] Eddy, Sean R. "Hidden markov models." Current opinion in structural biology 6.3 (1996): 361-365.

[10] Quattoni, Ariadna, Michael Collins, and Trevor Darrell. "Conditional random fields for object recognition." Advances in neural information processing systems 17 (2004).

[11] Chang, Yuan, et al. "Chinese named entity recognition method based on BERT." 2021 IEEE International Conference on Data Science and Computer Application (ICDSCA). IEEE, 2021.

[12] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[13] Deng, Na, Hao Fu, and Xu Chen. "Named entity recognition of traditional Chinese medicine patents based on BiLSTM-CRF." Wireless Communications and Mobile Computing 2021 (2021): 1-12.

[14] Alamoudi, Abrar, Amal Alomari, and Sarah Alwarthan. "A rule-based information extraction approach for extracting metadata from PDF books." ICIC Express Letters, Part B: Applications 12.2 (2021): 121-132.

[15] Jordan, Michael I., and Tom M. Mitchell. "Machine learning: Trends, perspectives, and prospects." Science 349.6245 (2015): 255-260.

[16] Gunaseelan, B., Supriya Mandal, and V. Rajagopalan. "Automatic extraction of segments from resumes using machine learning." 2020 IEEE 17th India Council International Conference (INDICON). IEEE, 2020.

[17] A. L. Xu, B. J. Liu and C. Y. Gu, "A recommendation system based on extreme gradient boosting classifier", *2018 10th International Conference on Modelling Identification and Control (ICMIC)*, pp. 1-5, 2018, July.

[18] 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.

[19] Joachims, Thorsten. Making large-scale SVM learning practical. No. 1998, 28. Technical report, 1998.

[20] Singh, Amit, et al. "PROSPECT: a system for screening candidates for recruitment." Proceedings of the 19th ACM international conference on Information and knowledge management. 2010.

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

[22] Bhoir, Nirmiti, et al. "Resume Parser using hybrid approach to enhance the efficiency of Automated Recruitment Processes." (2023)

[23] Wang, et al. " Chinese Electronic Resume Named Entity Recognition Based on BERT." China Sciencepaper 16.7 (2021).

[24] Song, Yu, et al. "POSBIOTM-NER: a machine learning approach for bio-named entity recognition." Korea 305 (2004): 350.