Construction of English Resume Corpus and Test with Pre-trained Language Models

Chengguang Gan¹ Tatsunori Mori¹ ¹Graduate School of Environment and Information Sciences Yokohama National University tmori@ynu.ac.jp

gan-chengguang-pw@ynu.jp

Abstract

Information extraction (IE) is an essential NLP task, especially for extracting information from resumes. This study aims to transform the information extraction task of resumes into a simple sentence classification task, using an English resume dataset and improving the classification rules to create a larger and more fine-grained classification dataset. The new dataset is used to test the performance of mainstream pre-training language models, and experiments are also conducted to compare the impact of different training set sizes on the accuracy of the resume dataset. The results show that the improved annotation rules and increased sample size of the dataset improve the accuracy of the original resume dataset.

1 Introduction

As artificial intelligence develops, using artificial intelligence instead of HR for resume screening has always been the focus of research. And the accuracy of resume screening depends on the precision of resume information extraction.Hence, it is crucial to improve the precision of resume extraction for the subsequent steps of various analyses performance of resumes. The previous study on resume information extraction tends to use the Bi-LSTM-CRF model for Name Entity Recognition(NER) of resume text[1].Although this method extracts the resume information (e.g. Personal information, Name, Address, Gender, Birth) with high accuracy, it also loses some original verbal expression information. For example, the description of one's future career goals, requires complete sentences that cannot be extracted by the NER method. As an AI system that scores the candidate's resume, the career object is also part of the score. In summary, sentences such as these should not be

ignored. Hence, in the prior study, the task of resume information extraction is transformed into a sentence classification task. Firstly, the various resume formats were converted into a uniform txt document. Then the sentences were classified after dividing them by sentence units. The classified sentences are used in the subsequent AI scoring system for resumes. The pilot study segmented and annotated 500 of the 15,000 original CVs from Kaggle.¹⁾ Five categories of tags were set: experience, knowledge, education, project and others²). The pilot study annotated resume dataset has problems, such as unclear classification label boundaries and fewer categories. Also, a dataset of 500 resumes with a total of 40,000 sentences in the tagging is sufficient for PLMs to fine-tune. If the dataset sample is increased, can the model's performance continue to improve.

To resolve all these problems, we improved the classification labels of resumes and used them to label a new resume classification dataset. To find out how many training samples can satisfy the fine-tune requirement of PLMs, we annotated 1000 resumes with a total of 78000 sentences. Furthermore, various experiments have been performed on the newly created resume dataset using the current mainstream PLMs.

2 **Related Work**

Since the last century, resume information extraction has been a critical applied research subfield in IE. In earlier studies, methods such as rule-based and dictionary matching were used to extract specific information from resumes[2]. HMM, and SVM methods extract information

¹⁾ https:

^{//}www.kaggle.com/datasets/oo7kartik/resume-text-batch

https://www.kaggle.com/datasets/chingkuangkam/ resume-text-classification-dataset





such as a person's name and phone number from resume information[3].Related Resume Corpus Construct study has an extensive resume corpus in Chinese[4].

3 Corpus Construction

3.1 Annotation Rule

We increased the number of categories from 5 to 7 in order to discriminate the various parts of the resume more carefully.As shown in Figure 1, the blue block on the left is the abbreviation of the seven classification labels, and on the right is the name of the resume section corresponding to the label.The full names of the seven labels are **Experience**, **Personal Information**, **Summary**, **Education**, **Qualifications**, **Skill**, and **Object**.The newly developed classification rules make it possible to have a clear attribution for each item in the resume. It will not cause the neglect of some sentences in the resume, as there are **other** labels in the prior study.

3.2 Annotation Tool

In order to label resume datasets faster and more accurately, we developed a simple annotation program based on Tkinter³⁾. The operation interface of the resume annotation tool. This tool automatically recognizes original resumes in pdf, Docx, and txt formats. It can also segment all the sen-

3) https://docs.python.org/ja/3/library/tkinter.html

NUMBER AND A STREAM							
	in Spectra	C. Diller M.	C Type Sue	C Spe Edu	C. 304.00	C Type Skill	1° Type
91-40-404	# type the	Cheelt	C Type Sum	C Type Edu	C 3pe 00	C Type Skill	C Type
	# tipe be	Cheeli	C Type Sum	C Type Edu	C 3pe 00	C Type Sill	C THP
baany	# her be	C hper8	C Tope Sum	C Tape Edu	C her 00	C Type Sel	C THE
14 years of superiouse in Andraing Damaness Breds, Caldering Depirements, Flamming, Cognising, Documenting, Damaness Provess Definition, Fa do Andrein, Tee Case Andrein and Antine inference and instance.	* her be	C hpe R	C Tope Sum	C Tape Edu	C Tax 00	C Type Sel	C Tree
Superior Innas Rawledge Fassor (Innernes, Berippy), 56/3-Commun. Bedikeur, Belix ad Rawlatering.	if here the	C here B	Clocks	C here Edu	C 1cm 00	C here fail	C Im
Spectre as follower Devignment Leis Cycle (EEC) and along follower Architecture, Speci Crimitel programming, Deripsing & Prototyping, Devinement, Education, David and Science and Production Description, Description, Construction, Desc	# Tope Dop	C hpe R	Chorsten	C Tupe Edu	C 104.00	C Tope Sell	C Tep
Reprinted with Estimal Datied Process (UD), Spile (Irea) Infrase Development, Tateofall Britanilary and UE.	if here here	Chevill	Clocks	C how Edu	C 1cm 00	C Tree Sel	C Inc
Interactive involving with deals anticology, form measures, creation of product hadding. Sprints as a part of the team with shared as and form review anticology for any part of the team with shared as and form review anticology. Development of the team with shared	of Tape Do	C Tape R	C hos ive	C Tupe Edu	C Table QC	C Tope Skill	C Tree
Apertics incology and approximent in Astropoles Researce Flanning (MP) and Cartanee Relationship Response (CRD).	if here for	Cheelb	Clocks	C Toy Mr.	C 100.00	C Tow Mil	C Ten
Reprinced in decign and development stages of P-Commerce reporting mode through advice parties to ensure the reporting table use to the department's standards. Which is turns will be used as a many for the decision-making process.	il Type Sap	C Type R	C Type Sue	C Type Edu	< 1pe 00	C Type Skill	C Tre
Hands on experience in various business applications to support and improve business operations including day to day administrations of DH AZ 400 Listian	il Type Sup	Cheele	C Type Sue	C Type Edu	C Type QC	C Type Skill	C TIP
Distancioning social on PACHY & compliance of 200 3 to 200 14 and R29ak (Boalth Inserance Perturbility and Accountability Act) 4004 and 8008 BM transmitters.	if type top	Cheve	C Type Sum	C Type Edu	C Type QC	C Type shill	C Typ
Dipersioner with Nodorara, Nedorard and connected, increments in EPAA 4002 312 fermate including (**) (**) 500, 500, 500, 500, 500,	P. Spectra	Cherry	C Type Sum	C Speciale	C Type QC	C type shill	C Type
Receipter with the writing of Nederary Part A. S. C and Ki and Nederald Nealth Instrume Polician. Personal Joiner Prefaction of P1 interaction a, Scilling experiptor with line and simulation in beach science with the read-understanding of P1 reduce. (One 1999 class from a	if type top	Chern	C type law	C type lide	C 1564 QC	C type thit	CTIP
Specieurs in Sevelupter, Administration, Configuration, Darlementation, Lightning and support experience with Salerforce OW plotfers.	4 type top	Count	C type law	C type lide	C type QC	C type thit	C Type
Billed in Tabless before for data vignalization through various charts such as bar charts. Line charts, contaction charts, ployt table, see The plots, ploy charts and peopled bubbles and use weiting a segmenter for comparison such as individual more. Hended more, and beal more	A type by	Chants	C type have	C type links	C type QC	C type thill	0.70
Received codecting from Sectowest Finning UDD and Joint Application beelayeest UND service with end-avers. SHI' and development tem for project methods, volathorships and outliner interviews.	A type by:	Count	C type law	Coperate	C type QC	C type thit	0.70
branetation includer: Bariners Prover Formentation, Dariners Prover Median, MIN's (2022)/MD1 and MIN'S (2014)20.	A tember	Chants	C Type Same	Chamble	C type QC	Coperated	0.70
Newloped for Gate Hagman. Sequence Hagman. Jura Flow Hagman. Process Flow Diagman and activity Hagman.	A type by	Chent	C Type Same	C Speciale	C type QC	C type thit	0.70
Inperienced developing Host Interface Inquirements. Host Interface (pecificentians, Hidvent Red nodego access), Montranes using Visits and For	P from the	Chent	C free law	C here links	C here oc	C from this	C Im
Reprinted in using #ES. Wills. IS office, Fower H & Swg.	A Spectra	Coperty	C Type Los	Chantele	(1gm QC	C type that	0.70
ferbel extensively with DM tools for dimensional mobiling, data migration, data transformation, data londing, data mobiling, metadata mobili, initian creation, ode mobiling. DR development and performance toning.	A top by	Cheen	C Type Sum	C Spe Ide	C 1ge QC	C Type Skill	CTIP
Experienced in understanding database structures and write SE/H29E.	P Deer De-	C heeft	C Tem Ser	C here bits	C here OC	C free Sel	C for
intenzively experienced in data analysis. MI techniques for loading high volumes of data and month structural flow of the data.	P See be	Cheell	C Tem Ser	C here lide	C here OC	C here fail	Clas
dept at wiring Data Reping Documents, Data Transformation Rulas, Reintaining Data Distinuary, Data Repution Documents and Datastana Repin	A bee te	Cheel	Cherles	C here Ide	C here OC	C free Sel	Che
Superioused in writing fost Flaux, Sest Cares, Courante Test Scripts and Baintain Reports for defest trading.	if here two	C here B	Chertin	C here lide	C here OC	C free Sel	Class
Speciment in test minution, idmitiging tests to be minuted, designing test minutes framewok, building minuted scripts, and type	if here fan	Cheell	Chersen	C here I de	C here 00	C how Sel	Class
Improved in Astronomics Period, System Period, New Astrophane Period (NC), Paretical Section/Hash Sec Period, Sec Period, Sec.	# her be	Cheen	Chersen	C how Edu	C Tape 00	C Tope Sal	C Tep
Interactivity represented of NE, and NEE. Todard with OE. format and imag able to real NE. false and use NP parsing table.	if here for	C here B	Clocks	C how Ide	C 1cm 00	C here fall	C Im

Figure 2: The operation interface of the resume annotation tool.

tences in the original resume according to a simple rulebased approach.Figure 2 shows the sample interface of the resume annotation tool. On the left are the sentences split by rule-based, and on the right are seven buttons that can be selected individually. After the sentence annotation of a whole resume is completed, a separate txt file will be automatically exported after closing the window, and the sentence annotation window for the next resume will be started automatically.

4 Experiments Set

In this section, we will perform various test experiments on the new-constructed resume dataset. First, we compared the performance of the BERT[5] model on the original resume corpus and the newly constructed resume dataset. Furthermore, four mainstream PLMs models are selected to test the resume dataset performance: BERT,ALBERT[6],RoBERTa[7], and T5[8]. For the fairness of the experiment, the size with the most similar parameters was chosen for each of the four models(BERT_{large}, ALBERT_{xxlarge}, RoBERTa_{large}, T5_{large}).The evaluation metrics for all experiments were F1-micro.The training set, validation set, and test set are randomly divided in the ratio of 7:1.5:1.5.And each experiment was performed three times to take the average of the results.

5 Result

5.1 Pre-train Models Test

As shown in Table 2, the new resume corpus ameliorated by 0.35% over the original dataset F1-score for the same BERT model. RoBERTa and T5 scores improved by 0.68%

Sample	10000	15000	20000	25000	30000	35000	40000	45000	50000	55000
Valid	83	83.7	84.9	84.9	85.6	86	86.1	86.6	85.9	85.9
Test	83.5	84.3	85.3	85.6	84.6	85.4	85.9	85.9	85.8	85.1

Table 1: The first row indicates the number of training sets. The following two rows indicate the F1-score of the validation set and test set corresponding to the number of training samples.

Model	F1-score
BERT* _{large} (baseline)	86.32
BERT _{large}	86.67
ALBERT _{large}	86.40
RoBERTa _{large}	87.00
T5 _{large}	87.35

Table 2: The first column * show accuracy of resume dataset before improvement.



Figure 3: F1-score of different training samples.

and 0.97% over baseline, respectively. The above results are also consistent with the ranking of the four PLMs in terms of their performance in various benchmark tests of NLP.

5.2 Sample Size Affects Experiment

In order to find out how many samples can bring out the maximum performance of the model, we divide the data set into training set 58000: validation set 10000: test set 10000. As shown in Table 1, the scores of the validation and test sets for different sample sizes. The model scores are tested from the 58000 training set, starting from 5000 and increasing the number of training samples every 5000.The highest score in the validation set is 86.6 when the training sample equals 45000. the highest score in the test set is 85.9 when the training sample equals 40000 and 45000.In order to visualize the relationship between the number of training samples and performance, we plotted the graphs (As Figure 3).It can be seen that as the number of training samples increases, the correctness of the



Figure 4: Fan chart of the percentage of each category of the resume corpus.

model rises. Finally, the model's performance reaches the highest point when the training samples are increased to 40,000. From the experimental results, for the PLMs, this resume corpus above 40,000 is sufficient for the model's maximum performance. The results also prove that the new resume corpus, which doubles the sample size, is significant compared to the original resum corpus.

6 Analysis

In the final section, we analyze the sample distribution of the constructed resume corpus.Figure 4 shows that the category with the most significant proportion in the resume corpus is **experience**, which accounts for half of the resume text. In addition, the three categories that account for the least in the resume corpus are **skill**, **object**, and **qualification**, which account for only 7%, 3%, and 1%.Conclusively, resume text is a very easy sample imbalance for experimental subjects. Thus, the resume corpus also vigorously tests the model's learning capability for categories with sparse samples in the training dataset. Hence, we plotted the conflation matrix of RoBERTa and T5 models. It is used to analyze the learning ability of the two models for sample-sparse categories in the dataset.

As shown in the figure 5, we can see the confusion matrix of RoBERTa and T5 models. First, the RoBE-RTa model is better for classifying **qualification** with the least number of samples. Secondly, the T5 model is slightly better than the RoBERTa model in terms of overall category classification results. The above results also demonstrate that our constructed resume corpus is highly unbalanced. However, if the model has strong performance, it can still learn the features of the corresponding category from very few samples.



T5 Confuse Matrix



Figure 5: Confuse Matrix of RoBERTa_{large} and $T5_{large}$ model in test set.

7 Conclusion

In this paper, we improve the classification labels of the original English resume corpus. Furthermore, it doubled the number of samples size. The final tests and analyses also show the reliability of the newly constructed resume corpus. In future work, we will explore how to solve the sample imbalance problem of the resume corpus. Make the model learn effectively even for small sample categories.

References

- S. Huang, L. I. Wei, and J. Zhang. Entity extraction method of resume information based on deep learning. Computer Engineering and Design, 2018.
- [2] R Mooney. Relational learning of pattern-match rules for information extraction. In Proceedings of the sixteenth national conference on artificial intelligence, volume 328, page 334, 1999.
- [3] Kun Yu, Gang Guan, and Ming Zhou. Resume information extraction with cascaded hybrid model. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL' 05), pages 499–506, 2005.
- [4] Yanyuan Su, Jian Zhang, and Jianhao Lu. The resume corpus: A large dataset for research in information extraction systems. In 2019 15th International Conference on Computational Intelligence and Security (CIS), pages 375–378. IEEE, 2019.
- [5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
- [6] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942, 2019.
- [7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.
- [8] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res., 21(140):1–67, 2020.