Refined Datasets for Competence-Level Classification and Resume & Job Matching

Boxin Zhao and **Jinho D. Choi**
Department of Computer Science
Emory University
Atlanta, GA, USA
{bzhao44, jinho.choi}@emory.edu

1 Introduction

The Human Resource (HR) departments of Companies and Institutions have long suffered from the overwhelming amount of resumes received for each and every job opening. More recently, multiple research have shown that advancements in Natural Language Processing (NLP) could effectively facilitate the resume screening process through classifying the resumes. In 2020, a group of researcher proposed two new tasks, Competence-Level Classification and Resume & Job Description Matching, viewing resume classification from the competence-level perspective instead of the category perspective. While their work suggested promising results, there are a few drawbacks in their work, potentially limiting the generalizability of their work: (a) the competence-level labeled resume dataset was noisy, (b) the job descriptions dataset lacked variety, (c) the context-aware model they developed is fairly complicated, and (d) the performance still has plenty of room for improvements.

This paper presents a follow-up research of (Li et al., 2020), improving the performance and reducing the complexity of deep-learning-based language models on the two proposed tasks through cleaning the labeled resume dataset and developing a new job description dataset. Our work suggests that with a refined dataset, a simple transformer model can perform just as well as the context-aware model. [1]

2 Related Works

Regarding resume classification, (Nasser et al., 2018) suggests promising results of using Convolutional Neural Network (Lecun et al., 1998) with Word Embedding Base Approach to classify resumes by domains. (Ali et al., 2022) provided a comprehensive study on the performance of doing resume classification using traditional Machine Learning and Natural Language Processing approaches, including K Nearest Neighbors (Murphy, 2012), Naive Bayes (Bishop, 2006), and many others. (M et al., 2023) pushes the study of classifying resumes by more detailed domains through the use of Ensemble Learning. However, by far the most relevant research in classifying resumes according to competence levels and matching resumes with job descriptions is still (Li et al., 2020), making use of the context-aware transformer models.

3 Approach

Task Definition In alignment with the work of (Li et al., 2020), we name Competence-Level Classification as `T1`, and Job Description Matching as `T2`. `T1` aims predicts the competence level of a resume, and `T2` aims to determine whether a resume and a job description make a good match.

Labeled Resume Dataset A CRC is a clinical research professional whose role is integral to initiating and managing clinical research studies. There are four levels of CRC positions, `CRC1-4`, with `CRC4` having the most expertise (Li et al., 2020). For `T1`, we use the same dataset used by (Li et al., 2020), which consists of 3425 resumes, each labeled with a human-expert-annotated CRC level. To evaluate the data, we randomly sampled 30 resumes and validated these resumes according to the CRC Hiring Guidelines provided by the Emory HR department. We found that 7 out of the 30 resumes we have sampled could be considered mislabeled. Therefore, we have conducted rounds of data revision, visiting 994 resumes in total, and fixed 332 out of the 994 resumes we have visited.

In addition to cleaning the data, we also added another 1500 of newly labeled resumes to the dataset, resulting in a total of 4925 labeled resumes used to train the `T1` model. Instead of re-splitting

[1] All our resources including the datasets and models are publicly available through our open source project: https://github.com/emorynlp/eclair-transformer

the data, we added the entire 1500 new resumes into the training set based on the previous split, by doing which we could better compare our results with the original work as well as enhancing model performance through training with more data.

Job Description Dataset Instead of using various distinct job descriptions, the T2 model in (Li et al., 2020) is trained on only 4 distinct job descriptions obtained from each CRC level, resulting in a lack of variety. To address this, we developed a job description dataset containing 710 entries of job descriptions through reverse engineering, meaning that each unique job description is developed from an actual resume from the resume dataset and therefore describes a unique CRC position. For each job description (except very few developed out of CRC3 and CRC4 resumes), we find in the labeled resume dataset 3 positive examples that match the description and 3 negative examples that do not. After the developement of this dataset, we now have 710 unique job descriptions with 4112 total positive/negative examples.

Because the job description dataset is newly developed, for T2, we will not split our data in the same way as the previous work did. Instead, we have conducted a simple 80-10-10 train-valid-test split.

Resume Parsing We put our resume dataset through an AI resume parsing tool by Rchilli[2], which parses a resume into more than 140 fields. We took 4 of the parsed fields, namely Qualification, Certification, Experience, and JobProfile, based on the advice from experienced recruiters of CRCs, concatenated the fields with the separator token <sep>, and used that as the content of resumes instead of the entire resume.

4 Experiments

Model Training In alignment with the work of (Li et al., 2020), we have selected the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) and Robustly Optimized BERT Pretraining Approach (RoBERTa) (Liu et al., 2019) as our models, as they have been reported to have decent performances on both T1 and T2 at a lower complexity in previous works.

For both tasks we selected `bert-large-cased` and `roberta-large` as the model. For T1, we used the concatenation of the fields parsed from the resume as input, and the competence-level labels as the expected model output. And for T2, we make use of the job description dataset that we have newly developed. For each job description, we take its positive and negative example resumes, and then concatenate the job description at the end of each fields concatenation extracted from the resume using the separator <sep>, and use that as the input for training. The expected output label is "YES" if the resume is a positive example, and "NO" if it is a negative example.

Results Labeling accuracy is used as the evaluation metric for all our experiments. Table 1 shows the labeling accuracy on the test set of our model and the model of the previous work, for T1 and T2, respectively.

	ACC
(Li et al., 2020)	73.26 ($\pm$ 0.16)
Ours-BERT	74.83 ($\pm$ 0.53)
Ours-RoBERTa	75.65 ($\pm$ 0.92)

Table 1: Model Accuracies for T1

From the results we can see that for T1, our naive RoBERTa model successfully beat the context-aware model that was previously developed. For T2, our model have achieved an accuracy of 74.98% ($\pm$ 2.08). Although the accuracy is lower than what was reported in the previous work, our model was trained and tested on a newly developed job description dataset with 710 distinct job descriptions, each with 6 positive/negative examples, whereas the context-aware model was only trained and tested on 4 job descriptions.

5 Conclusion

Compared to the previous work conducted by (Li et al., 2020), we have corrected a considerable amount of competence-level labels and developed a new job description dataset to enhance model performance in both tasks proposed in the previous work. Our work has proven that data quality could considerably influence transformer models' performance on Competence-Level Prediction. In the future, we will explore leveraging the power of more recent, better performing models to improve the performance.

[2]https://www.rchilli.com/

References

Irfan Ali, Nimra Mughal, Zahid Hussain Khand, Javed Ahmed, and Ghulam Mujtaba. 2022. Resume classification system using natural language processing and machine learning techniques. *Mehran University Research Journal Of Engineering & Technology*, 41(1):65–79.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer, New York.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Changmao Li, Elaine Fisher, Rebecca Thomas, Steve Pittard, Vicki Hertzberg, and Jinho D. Choi. 2020. Competence-level prediction and resume & job description matching using context-aware transformer models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8456–8466, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Spoorthi M, Indu Priya B, Meghana Kuppala, Vaishnavi Sunilkumar Karpe, and Divya Dharavath. 2023. Automated resume classification system using ensemble learning. In *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 1782–1785.

Kevin P. Murphy. 2012. *Machine Learning: A Probabilistic Perspective*, 1st edition. The MIT Press, Cambridge, Massachusetts.

Shabna Nasser, C Sreejith, and M Irshad. 2018. Convolutional neural network with word embedding based approach for resume classification. In *2018 International Conference on Emerging Trends and Innovations In Engineering And Technological Research (ICETIETR)*, pages 1–6.