Research Report - Scholar Profiling using Resumes

Vedaant Jain

Contributing authors: vvjain3@illinois.edu;

1 Introduction to the Problem

Scholar profiling plays a crucial role in summarizing the achievements and contributions of research scholars, offering organizations a means to efficiently evaluate scholars for various opportunities. Resumes and Curriculum Vitae (CVs) serve as vital repositories of information about scholars, making the extraction of relevant data from these documents essential for enhancing the accuracy of academic scholar profiling.

While certain techniques have been developed to extract information from visual documents and have been successfully applied to resumes[1], their effectiveness is limited when it comes to extracting information from academic resumes. Unlike typical resumes, academic resumes are often multi-page documents, with content spanning across multiple pages. Extracting and organizing information into discrete sections, such as publications or work experience, pose a challenge that current resume parsers fail to address adequately. Moreover, academic resumes often employ similar terminology across various sections, such as education, employment, and scholarly activities, which can lead to confusion for existing models that heavily rely on pre-trained embeddings to extract relevant information[2]. Additionally, academic resumes tend to have a similar structure, that most current parsers do not take advantage of.

2 Early Approaches and Challenges

Initially, we attempted to fine-tune models like LayoutLMV2 for resume information extraction. However, we encountered limitations, including the models' inability to effectively extract information from longer textual sections, such as the work experience section. Additionally, the models struggled with confusion between education and employment labels due to recurring words.

Initially, we attempted to use whitespace identification with EasyOCR to section resumes. However, this approach proved inadequate due to several limitations. It struggled with multi-page resumes, often incorrectly splitting or merging sections at page breaks and whitespace. Additionally, the method faced challenges with single-column

academic resumes, particularly when horizontal whitespace caused erroneous section divisions.

Working on the assumption that academic resumes are single-columned, we built our solution to section a resume using title detection and then carrying out information extraction.

3 The proposed solution

In this report, we introduce an approach to structurally segmenting a resume by utilizing the widely adopted title-content style commonly found in academic resumes. By leveraging this structure, we enable more precise information extraction from specific sections, thereby mitigating potential labeling confusion within the model. For instance, without proper sectioning, the model may mistakenly label words within the employment section as the university where a scholar obtained their education. Through our framework, such mis-classifications are minimized, ensuring accurate information retrieval.

3.1 Sectioning a resume

To accurately section resumes, we leverage title detection models, particularly the Document-Image-Transformer (DiT) introduced by Li et al. (2022)[3]. DiT has demonstrated excellent performance in document layout analysis, achieving a score of 0.9491 on the PubLayNet dataset.

To evaluate DiT's performance specifically for title detection in resumes, we manually labeled titles on 25 resume pages using Label-Studio in the COCO format for image segmentation. This ground truth dataset allows us to assess the accuracy of DiT in identifying titles. However, we encountered a challenge where DiT often recognized sub-headings as main headings, which required further refinement.

To address this issue, we utilized the semantic similarity between academic resume headings and sub-headings. We compiled a list of 25 common titles found in academic resumes. By comparing the word2vec embeddings of these common titles with the word embedding of the detected titles, we could differentiate between sub-headings and main headings. To measure similarity, we utilized the cosine similarity metric, defined as follows:

$$\cos(\mathbf{t}, \mathbf{e}) = \frac{\mathbf{te}}{\|\mathbf{t}\|\|\mathbf{e}\|} = \frac{\sum_{i=1}^{n} \mathbf{t}_i \mathbf{e}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{t}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{e}_i)^2}} \tag{1}$$

Additionally, we compared the bounding boxes of the detected titles with the ground truth labels, employing the Intersection-over-Union (IoU) threshold. If a detected title had an IoU greater than 0.45 with a ground truth label, we considered it a true positive. **By using this evaluation metric, we computed a precision of 0.88 and a recall of 0.90, indicating the effectiveness of our title detection approach.** The code for this can be found here.

3.2 Extracting information from a section's text

After sectioning a resume page image to sections, we run OCR on the section to get the text present. At this stage we have trained models to extract information from education, employment, and publication sections.

We only use textual features to extract information from each section because we already know what kind of section we are getting information from, and so in order to use a lighter model, we could get good performance if we ignore the visual features of the specific sections.

We use a sequence tagging approach to get information from the text of the sections. Specifically, we use a BIO(Beginning, Inside, Out) tagging appraoch. We created a dataset using text from education, employment, and publication section and labelled data in the CONLL format. We extract the following information for the following sections:

Section	Information Extracted
Education	Degree, University, and Thesis
Employment	Designation, and Employer
publication	Authors, Title, Journal

After dividing a resume page image into sections, we apply Optical Character Recognition (OCR) to extract the corresponding text from each section. At this stage, we have trained models specifically designed to extract information from education, employment, and publication sections.

As we already know the type of section we are extracting information from, we solely rely on textual features during the information extraction process. This allows us to employ lighter models, as the visual features of the sections can be disregarded without sacrificing performance.

For information extraction, we utilize a sequence tagging approach, specifically employing the BIO (Beginning, Inside, Outside) tagging scheme. To train our models, we construct a dataset comprising text from education, employment, and publication sections, which is labeled in the CONLL format. This labeled dataset serves as the basis for training our models to accurately extract the desired information.

We utilize BERT embeddings for sequence tagging and perform token classifications based on the tokenized output of BERT. By adding linear layers on top of BERT, we enable token classification using the Huggingface library's implementation of BertForTokenClassification. Each section type (education, employment, publication) has its own dedicated model, which is fine-tuned on the respective dataset. With these models, we can effectively extract information from the relevant sections of the resumes.

To determine the section type (education, employment, or publication), we leverage the previously extracted title information. By using the extracted titles, we are able to accurately assign the corresponding sections, thereby facilitating the information extraction process.

In summary, our approach combines BERT embeddings, token classification using BertForTokenClassification, and the BIO tagging scheme for sequence tagging. This

enables us to train specialized models for information extraction from education, employment, and publication sections.

3.2.1 Results of Information Extraction

We report the accuracy for information extraction from each section as follows:

Section	Accuracy
Education	0.84
Employment	0.87
publication	0.85

4 Assessment, Reflection, Future Work

In conclusion, our research report introduces a framework for information extraction from academic resumes, addressing limitations and improving scholar profiling. By combining structural sectioning, sequence tagging, and multi-page resume handling, our methodology offers a comprehensive solution for accurate information extraction.

One notable advantage is the seamless handling of multi-page resumes, capturing essential details across multiple pages. This enhances the thoroughness of scholar analysis, resulting in more comprehensive profiles.

The creation of a labelled dataset with 30+ academic resumes, totaling more than 250 resume pages, strengthens our models for sectioning and information extraction, ensuring reliability and adaptability in real-world scenarios.

By adopting a modular approach, where sectioning and information extraction tasks are separate, our methodology provides greater transparency and interpretability. Researchers can better understand and improve the models by analyzing the results of each task individually. This transparency facilitates ongoing research and development, enabling the refinement of the models based on insights gained from the sectioning and information extraction processes.

There are several limitations to the approach as well. The method to separate headings and sub-headings is not very generalizable, as a new heading might not be recognized, especially if there is no embedding for it in word2vec. There is also not much research gone into differentiating headings and sub-headings, which presents a great area to do further research in.

In addition, we could have experimented with various other models for text-based information extraction. One example is using a Bi-LSTM CRF with FLAIR word embeddings for sequence tagging.[4][5]

Another aspect that can be improved on is OCR-post correction, this was in progress but not completed satisfactorily before the semester end. For resumes that are originally in pdf format, we compare the OCR reading and original text extracted from the pdf, and correct any errors in words using the Levenshtein distance similarity between words.

Reflecting on this research journey, I have gained valuable insights into various NLP tasks such as sequence tagging, named entity recognition (NER), and document

understanding. I have also deepened my understanding of transformers and their significance in diverse applications. Through reading and evaluating research papers, I have honed my skills in paper analysis, enhancing my ability to comprehend and interpret academic literature.

Looking ahead, I recognize the importance of more comprehensive reflections for each week's work and utilizing the provided spreadsheet not only to track progress but also to critically evaluate methodologies and identify areas for improvement. For instance, I now realize that earlier exploration into the lack of research on title and sub-title differentiation could have led to a novel and valuable contribution to this project, despite potentially causing some delays.

Acknowledgement: I extend my sincere gratitude to Professor Kevin C. Chang for his invaluable guidance and mentorship throughout the semester. His expertise and support have greatly enhanced my research skills, enabling me to grow as a researcher. The weekly meetings have been instrumental in discussing solutions, addressing challenges, and receiving expert advice. I am truly grateful for the opportunity to learn and collaborate under his guidance.

Overall, this research endeavor has been a transformative learning experience, equipping me with a deeper understanding of NLP tasks, model development, and research processes. I am excited to continue my research journey, applying the knowledge and skills gained from this project to contribute further to the field of natural language processing.

References

[1] Cheng, Z., Zhang, P., Li, C., Liang, Q., Xu, Y., Li, P., Pu, S., Niu, Y., Wu, F.: TRIE++: Towards End-to-End Information Extraction from Visually Rich Documents (2022)

[2] Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., Zhang, M., Zhou, L.: LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding (2022)

[3] Li, J., Xu, Y., Lv, T., Cui, L., Zhang, C., Wei, F.: DiT: Self-supervised Pre-training for Document Image Transformer (2022)

[4] Alzaidy, R., Caragea, C., Giles, C.L.: Bi-lstm-crf sequence labeling for keyphrase extraction from scholarly documents. In: The World Wide Web Conference. WWW '19, pp. 2551–2557. Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3308558.3313642 . https://doi-org.proxy2.library.illinois.edu/10.1145/3308558.3313642

[5] He, H., Choi, J.D.: Establishing Strong Baselines for the New Decade: Sequence Tagging, Syntactic and Semantic Parsing with BERT (2020)