Information extraction from and summarization of Scientific Curriculum vitae

Pavan Balaji Kumar
300169572
pkuma079@uottawa.ca
CSI 6900 – Graduate Project
Supervisors: Dr. Diana Inkpen
School of Electrical Engineering and Computer Science (EECS)
Faculty of Engineering
University of Ottawa

Abstract

With the improvement in text parsing and analysis using natural language processing techniques, manual curriculum vitae evaluation has become a thing of the past. Though parsing and extracting information from resumes and curriculum vitae has become easier in most of the industries, it is still a difficult task in the research community. Parsing through a scientific document which is unstructured and does not have a common template to be followed in a challenging task. Since most of the topics in the curriculum vitae of researchers are not the same as the other disciplines traditional resume or CV parsers do not work well here. To parse and collect data from scientific curriculum vitae we propose a new system using a natural language processing technique, Name entity recognition. By performing name entity recognition, we first label all the entities present in the curriculum vitae with different tags like person, awards, events etc. Once labeling is done, we use that information to scrap more information related to the author and store is as the dataset. Once the information about a researcher is stored in a dataset it is then used to provide a summary of that researcher using Data2Text method. This will help researchers decide on possible collaborations and evaluate peers in their filed.

Keywords: Natural language processing, name entity recognition, Curriculum vitae, Parsing, Data2Text.

1. Introduction

The Curriculum Vitae's (CV) of researchers can be considered one of the important documents in the scientific community. The document contains information about the research projects, publications, awards, and conferences undertaken by them. It also provides plethora of other information such as organisation, supervisors, research standard and may more [1]. These are just some of the metrics which are used to evaluate a researcher performance in a particular discipline and academic success [2]. The CVs of a researchers also works as job search resource and are evaluated carefully before recruiting or collaborating with one another [2]. Though there is a standard for what information to include in the CV, there is no clear structure followed by the researchers, which makes the perusing and evaluating CV's a tedious process and a lot of research on automating this process is being done.

In this project we aim to automate the information extraction and evaluation of Scientific Curriculum vitae to increase the efficiency CV evaluation phase and help researcher's find suitable partners or assistant for their research work. This will be accomplished using various natural language processing techniques. Methods like Name entity recognition (NER will be used for extraction of required information from the CVs. Theses methods are trained to identify all the entities of the same type in a text or document efficiently [3]. Pre-trained deep learning models such as BERT, RoBERTa [4] will be used for the NER task to identify entities in the CVs like name publications, awards, conferences etc. The labels identified are then used to collect information from sources like google scholar and Web of science. This extracted information is then processed into a Dataset for downstream tasks such as summary of an individual CV. Summarization task is done using a popular natural language method called Data2Text, where a summary or a description is formed using structured data. Once a curriculum vitae of a researcher is summarised, it can be used to decide on possible research collaborations and selecting good candidates for certain scholarships and fellowships.

2. Related works

In this section we will discuss some of the research works already present in the field of resume parsing and name entity recognition. In the paper [5] various techniques for Name entity recognition are discussed. It gives a clear picture of how a

NER system works and also gives information on how a custom NER pipeline can be built and also briefly discusses various pre-built NER systems. In this project we have used one such pre-built NER from Spacy.

Design challenges that are to be considered while building a Name entity recognition system like how to represent text chunks, what kind of algorithm to use for inference, what other dependencies we can use for inference and what other resources will aid the betterment of a NER system [6]. It gave an clear understanding of how to build an name entity recognition system considering everything possible. A machine learning based system combining convolutional neural networks (CNN), Bidirectional long short term memory model (Bi – LSTM) and conditional random fields (CRF) is proposed in [7] to parse resumes and curriculum vitae. In this paper the CNN model is used to classify the segments in the Resume after it has been converted to text. Bi-LSTM and CRF is used to label is used to labels the segments text which is converted into JSON datapoints. This paper gives a clear idea of how deep learning can be utilized in the field of Name entity recognition.

Visual Question answering system and various methods to implement them were discussed in [8]. These AQ systems typical have two modules, a computer vision module to process images and natural language processing module to process textual data. By using question based on the image can be answered. Summarisation techniques used in this method for answering are of a lot of types catering to specific needs. Another system, that is custom model built for name entity recognition is SANE 2.0, which is primarily used for name entity typing. It uses the Wikipedia database to try and get a more detailed entity tagging for the tegs identified by traditional NER models like Stanford, Spacy etc. It works by building static databases from Wikipedia and using it to train the models. Building such a database and following a similar method will give better results while parsing scientific resumes too.

A deep learning model trained in semi supervised learning method which is capable of specifically parsing through education section of resume or curriculum vitae is proposed in the paper [10]. It uses the BILOU method for tagging entities in the education part of the resume. Highlight and tagging specific elements in part of resume is a very useful technique.

3. System architecture

In this section we will look at how the system function and what processes are carried from the time a Scientific curriculum vitae is given as an input and a summary is created for it.

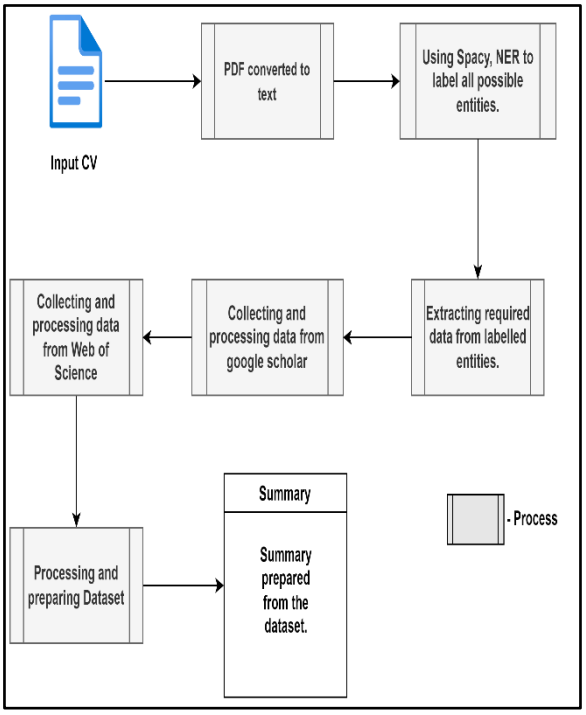

Figure 1. System architecture

Figure 1 shows the system architecture or the flow diagram of our project. The process flow starts when the resume is converted to text file. These text data is then passed to the pre-trained name entity recognition model and the model labels all possible entities in the document. Based on the labels provided, the textual data is processed, and data is extracted. This data is stored as a csv file. Using the name of the researcher additional information about him/her is retrieved from google scholar and web of science portal automatically and certain information are stored as a csv. Once all the required data is collected all the three files are combined as one using the name of the researcher as the key. Once the dataset is created, Data2Text method is used to form summary from the data points of each researcher, and this is available for the user to view and make decisions. The system has four exclusive modules which we will discuss about these modules in detail in the following session.

4. Module Description

The system proposed has a total of four modules which are,

I. PDF conversion
II. Name entity recognition
III. External Data Extraction
IV. Summarization

each modules serves a particular purpose and are pivotal for the systems functionality. Let us take a in-depth look at the modules.

4.1. PDF Conversion

This module focuses on converting the resumes or curriculum vitae in pdf format to plain text format. This is an important step

because working with pdf files directly can be a tedious task as they are structured in a way that information present in the pdf are represented as boxes or regions containing information and there is no direct way to access this data. Only specific packages or software are able to reproduce the pdf to plain text. Converting any pdf file to plain text is import for this project as it is required by the pre-trained NER model that the data be in text format.

To convert the pdf to plain text we use the PyMupdf python package. PyMuPdf is a python package which has all the functionalities that an normal pdf processing software has and also other functionalities that will help the developer to handle data in this files [11]. We open the files using the packages open() method which creates an document object. This document object contains the pdf file from which the text can be extracted using extract_text custom function. The extracted text is then processed using custom function containing regular expression written to remove unwanted punctuations, newlines etc. This processed text is stored as strings on which Name entity recognition is performed.

4.2.Name entity recognition

Name entity recognition is the second step of the proposed system. The text which was extracted and processed in the second step will be tagged using pre-trained model. In this project we will tag entities in the document using the Spacy pre-trained NER model. Spacy is a open source natural language processing framework providing high level api to perform various natural language processing tasks like sentiment analysis, name entity recognition, paraphrase detection, part of speech tagging etc [12]. It also provides lot for api for text preprocessing tasks like stemming, lemmatization, tokenisation and many more.

Figure 2. Sample NER output

For our project we will using the spacy transformer pretrained model “en_core_web_trf” to perform name entity recognition on the scientific curriculum vitae

of various researchers. The model give labels like PERSON, ORG, MONEY, GPE, LOC, DATE, CARDINAL etc. Figure 2 shows the sample output for NER using the spacy pre-trained model. The labels given are then used to collect relevant data and is stored as csv file.

4.3.External Data Extraction

Using the name of researcher from the CV, additional data with respect to publishing was obtained from external sources like Google scholar and Web of science. These data was automatically collected using python scripts, processed and then stored as csv files. First data collection is done from google scholar. A python package called scholarly is used to collect this data[13]. This data is present in JSON format. The required data from JSON is extracted and then stored in a csv file. This is done for all the CVs in the original dataset or just an single CV. If the information can not be found on google scholar then the researcher profile is searched on web of science portal.

Unlike google scholar there is no package or ready to use free api for web of science data information but it can access using paid api service. For this project purpose I decided to use the University of Ottawa employee access for web of science to acquire this data. Instead of manually downloading all the required data I wrote a selenium script which is capable automatically searching and downloading the researcher data from the web of science portal given the access credential of a university of Ottawa employee. Once the data for the given researchers are downloaded this is then processed and the same information extracted from google scholar is also extracted from this data. Once all the required data is extracted this two datasets are joined together. This is finally merged with the CV dataset to form the final dataset on which will be used for summarization task. The final dataset consists of the features like Name, Department, Awards, H-index, No of publication, total number of citation, number of years the researcher was active and no of conference and event the researcher has attend or presented in.

4.4.Summarization

For the summarization of the curriculum vitae based on the Dataset created, Template based Data2Text method was used [14]. The data points are fitted into this template. If some of the data are missing then that information is conveyed accordingly.

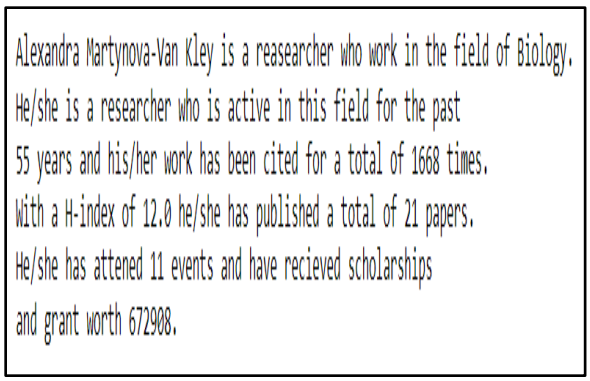

Figure 3. Sample summarization Output

Figure 3 show the sample output summary created using a template and the dataset created at the end of three modules. This summary can be used to perform evaluation of a CV and make any required decisions.

5. Discussion and Future work

This project as mentioned before aims to increase the efficiency of the scientific curriculum vitae evaluation. Though the system proposed does a good job achieving the desired goals and provide a good summary of the CV aiding in the evaluation process a lot more can be done from this point forward. For example with some of the CVs the pre-trained models are still struggling to find the appropriate entities and label for the entities. To overcome this custom MER model can be built from scratch by collecting enough CVs and preparing a dataset for NER model training. Though this is a time consuming and tedious process it will ensure that we identify what we want with correct labels. Another improve that can made to this prototype is that the use of a paid api for web of science as it drastically reduces the time to get data for a researcher compared to the script using selenium. This project can be extended into a full application which can provided information about researchers are also possible suggestion about them.

6. Conclusion

In this project we have built a prototype system which can parse a scientific curriculum vitae and extract information. We also incorporated additional external information about a researcher based on the information we gathered about them from the CV and created a static dataset consisting of various features mentioned earlier. The summary created using these can help the user understand about the researcher and take decision regarding collaborations, grants, awards and possible mentorship etc. If this is developed as an fully functional application as mentioned in the previous section, this can become a useful tool in the field of research and education.

References

[1] Cañibano, Carolina & Bozeman, Barry. (2009). Curriculum vitae method in science policy and research evaluation: The state-of-the-art. Research

Evaluation - RES EVALUAT 18. 86-94. 10.3152/095820209X441754.

[2] Fischer, J., Ritchie, E. G., & Hanspach, J. (2012). Academia's obsession with quantity. Trends in ecology & evolution, 27(9), 473-474.

[3] Fareri, S., Melluso, N., Chiarello, F., & Fantoni, G. (2021) SkillNER: Mining and mapping soft skills from any text. Expert Systems with Applications, 184 doi:10.1016/j.eswa.2021.115545

[4] Yang, S., Yoo, S., & Jeong, O. (2020). DeNERT-KG: Named entity and relation extraction model using DQN, knowledge graph, and BERT. Applied Sciences (Switzerland), 10(18) doi:10.3390/APP10186429

[5] Hemlata Shelar, Gagandeep Kaur, Neha Heda & Poorva Agrawal (2020) Named Entity Recognition Approaches and Their Comparison for Custom NER Model, Science & Technology Libraries, 39:3, 324-337, DOI: 10.1080/0194262X.2020.1759479

[6] Lev Ratinov and Dan Roth. 2009. Design Challenges and Misconceptions in Named Entity Recognition. In Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009), pages 147–155, Boulder, Colorado. Association for Computational Linguistics.

[7] C. H. Ayishathahira, C. Sreejith and C. Raseek, "Combination of Neural Networks and Conditional Random Fields for Efficient Resume Parsing," 2018 International CET Conference on Control, Communication, and Computing (IC4), 2018, pp. 388-393, doi: 10.1109/CETIC4.2018.8530883.

[8] Modi, Shivangi & Pandya, Dhatri. (2019). VQAR: Review on Information Retrieval Techniques based on Computer Vision and Natural Language Processing. 137-144. 10.1109/ICCMC.2019.8819803.

[9] Anurag Lal, Ravindranath Chowdary C., SANE 2.0: System for fine grained named entity typing on textual data, Engineering Applications of Artificial Intelligence, Volume 84, 2019, Pages 11-17, ISSN 0952-1976,https://doi.org/10.1016/j.engappai.2019.05.007.

[10] Gaur, B., Saluja, G.S., Sivakumar, H.B. et al. Semi-supervised deep learning based named entity recognition model to parse education section of resumes. Neural Comput & Applic 33, 5705–5718 (2021). https://doi.org/10.1007/s00521-020-05351-2

[11] https://pymupdf.readthedocs.io/en/latest/

[12] https://spacy.io/usage/v3-2

[13]https://github.com/scholarly-python-package/scholarly

[14]https://www.microsoft.com/en-us/research/project/data2text-automated-text-generation-from-structured-data/