ISMT S-117 Text Analytics and Natural Language Processing (2020 Summer)

Final Project Proposal

Resume Screening System Using BERT

by Micheal Choi Kwong Chak

Table of Contents

[Introduction 3](#_Toc46776501)

[Problem Statement 3](#_Toc46776502)

[Project Objective 3](#_Toc46776503)

[Description of Dataset 3](#_Toc46776504)

[Resumes and Exploratory Analysis Completed 3](#_Toc46776505)

[Job Descriptions and Exploratory Analysis Completed 4](#_Toc46776506)

[Proposed Additional Exploratory Analysis 6](#_Toc46776507)

[Description of Methodology Used 7](#_Toc46776508)

[Building the Training / Testing Corpus 7](#_Toc46776509)

[Models Setup 7](#_Toc46776510)

[Tools Used 7](#_Toc46776511)

[Deployment Strategy 8](#_Toc46776512)

[References 9](#_Toc46776513)

# Introduction

In the labor market, resumes and job descriptions are the documents to describe the applicants and the jobs they applied, respectively. These are natural candidates for natural language processing and enable both the HR professionals to find their best candidates for the jobs, and vice versa. With 98% of resumes that needs to be rejected (Elmers, n.d.) before a candidate worth interviewing is found, it is a great productivity improvement to build a recommendation engine for an job application tracking system.

# Problem Statement

Traditional resume parsers often focus on key word matching and sometimes supplemented with results of NER (Named Entity Recognition) (Wu, 2019), or one step forward, compare document similarities using fixed document representations like doc2vec or GloVe (Raman, 2019). However, the content of both resumes and job descriptions are 1) context sensitive, for example, two jobs named “software engineer” can mean very different things, and 2) domain knowledge required to succeed in the job are not necessarily described even in the job description. Optimally, to build a recommendation engine one needs to provide resumes labelled as suitable for interview for the job. However, as the job descriptions of the same job evolves over time, and that the goal of the recommendation engine is to save human labor, 3) such data is likely unavailable, at least for the initial stage of the system build up.

# Project Objective

The project targets to resolve problem 1) by using BERT (Devlin, et al., 2019) to learn the word representation under the context it is given so as to provide a better baseline for similarity calculation.

For problem 3), while resumes matching particular job descriptions are not readily available, an insight that the people holding similar positions will have their resume written in a way that would be worthwhile for an interview. With these resumes labeled as likely candidates in the training data into BERT for a particular job description, it is expected to have an improvement on the accuracy.

With the model learning other resumes and job descriptions documents, the resultant model will also contain knowledge to the domain and not just the job description in question.

# Description of Dataset

In this project we are targeting to limit both resumes and job descriptions to IT related jobs to limited the context and vocabs to be learned by the model. And so I have selected the following data sets under this criterion:

## Resumes and Exploratory Analysis Completed

* Resume Dataset from Avani Siddhapura (Siddhapura, 2020) <https://www.kaggle.com/avanisiddhapura27/resume-dataset>
  + This data set has 14800 IT related resumes with past job experiences, education records, certificates and skills parsed into separate sections and this is optimal for selecting target description for particular jobs title for labelling as mentioned above

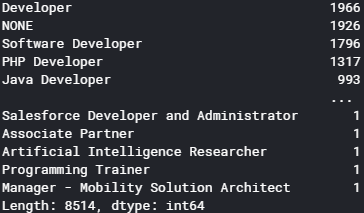


Figure 1: Counts of Job Title in Avani Resume Dataset

* + As this is segregated into sections, each job description entry in the past experience of the resume is relatively short with many null values which may or may not be a good thing in terms of training. It is possible that the sections need to be aggregated back for training.

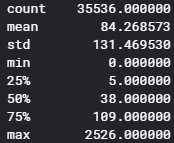


Figure 2: Distribution of # tokens for each job descriptions in Avani Resume Dataset

* + All resumes in this dataset is from India which could create a bias in the training but that is common to all resume datasets available, so we will have to keep this in mind on result interpretation. As more relevant data is generated from a deployed system, the model can be retrained for data that conform to what the population is like for the particular company.
* Resume Datasets from Amita Dhainje (Dhainje, 2019) and Maitri Palan (Palan, 2019)
  + <https://www.kaggle.com/dhainjeamita/resumedataset> (Size: 169, Type: mainly tech)
  + <https://www.kaggle.com/maitrip/resumes> (Size: 1219, w/ 104 under IT category)
  + These are classic resume datasets with only 2 fields, category and the resume itself. These resumes have all the fields aggregated together but with the original structure intact. These can potentially be used as the testing set. However, many of them do not have the job title identified which could make it hard to evaluate the correctness of the classification

## Job Descriptions and Exploratory Analysis Completed

* Online Job Postings – Armenian Online Job Posts from 2004 – 2015 (Hab, 2017) <https://www.kaggle.com/madhab/jobposts>
  + This dataset has 19000 online job posting on Armenian human resource portal, with 3759 IT jobs covering all areas.

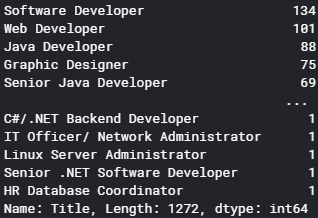


Figure 3: Job Titles in Armenian Job Posting Dataset

* + The data is a bit dated but 50% of all IT jobs are after 2010
  + The average # of tokens for each post is around 300 which is suitable for BERT sentence pair classification (max 512)

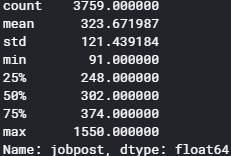


Figure 4: # Token Distribution for Armenian Job Posting Dataset

* Australian Job Listings Data from Seek (PromptCloud, 2020) & Gumtree (JobsPikr, 2019) Job Board in 2018(JobsPikr, 2019) (JobsPikr, 2019) <https://www.kaggle.com/PromptCloudHQ/australian-job-listings-data-from-seek-job-board> <https://www.kaggle.com/JobsPikrHQ/australian-job-listings-data-from-gumtree>
  + The dataset is extracted from seek.com and GumTree in Australia in 2018 containing 19,000 & 8,000 records respectively. Unfortunately, For Seek.com only 438 records are in the IT category AND with job description populated while GumTree has 277 IT related entries
  + These datasets are more to-date and with an Australian background.
  + However, the job titles are more diverse which can be a disadvantage to our matching

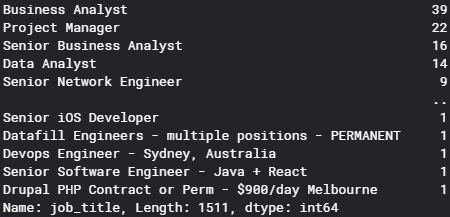


Figure SEQ Figure \\* ARABIC 5: Job Titles from Australian Seek Job Description Dataset

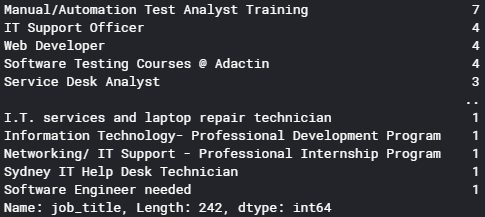


Figure SEQ Figure \\* ARABIC 6: Job Titles from Australian GumTree Job Description Dataset

* + On both sizes, the job description length is around 250 on average which is suitable for BERT training.
  + These datasets can be used to supplement the Armenian dataset or be used as testing sets as they are newer.

# Proposed Additional Exploratory Analysis

* We can use GloVe and pretrained BERT embeddings and to see how particular jobs (e.g.: Java Developers) are positioned in the vector space, summarized by t-SNE. This also provide a baseline to how BERT performs after the fine-tuned BERT embeddings become available.
* If there is time, we can also use k-means clustering to see if similar job titles can be identified by the clustering
* We can use LDA to see how particular jobs are factored in the topic model and whether they correspond

# Description of Methodology Used

## Building the Training / Testing Corpus

As we do not have a labeled corpus during the initial training of the model, similar job titles embedded using the pretrained model to calculate the embedding for similarity. Similarities over a threshold are labeled as a match for the training set. As more resume-job description matches are generated during the course of daily HR activities after the model is rolled out, a better set of data can be used to train the model.

As mentioned in the previous sections, the Armenian resume data set, where each job description, written by the current holder of the position, is matched with the job title, is used to train the model, matching randomly with the job descriptions to build the training set. The other resume data sets with full resumes are used as the testing set. Whether the other sections of the original resume like education history or certification will also be tested.

Initially we will only use IT related jobs for training. If time allows, we will extend this into all job categories and compare the results.

## Models Setup

In this project, we are going to build two variants of the implementation using BERT, including:

* Use sentence pair classification model to classify whether a candidate of a resume is worthy of an interview
* Fine-tune a BERT model to generate customized embeddings for cosine similarity computation

One of the General Language Understanding Evaluation Benchmark (GLUE) tasks tested on the original BERT model is to find similarity on the STS-B corpus. While the original corpus is much shorter than job description or a resume, in essence it is still a similarity classification of two sentences so it is very suitable to deal with this kind of problems. Supposedly, the domain knowledge not really present in the particular job description is also captured in the model – so it would supposedly provide improved prediction capabilities

On the other hand, the task above should not be different from simply fine-tuning a pretrain-ed model and embed documents with the model, as all decision information should come from the embeddings. Having the embedding of a document allows us to perform more tasks like visualization on document relationships.

## Tools Used

In this project we are going to do both and compare on the accuracy and usability of these implementations. In either case we are going to use the Simple Transformer library CITATION Raj20 \l 1033 (Rajapakse, 2020) <https://github.com/ThilinaRajapakse/simpletransformers#simple-transformers> which is a wrapper for the HuggingFace CITATION Hug \l 1033 (Huggingface, n.d.) implementation on PyTorch.

On visualization we are going to use a mix of matplotlib and Weights and Bias to monitor training progress.

# Deployment Strategy

This recommendation engine, after trained, can accept new resume-job pair for classification of worthiness of candidate for interviews. Initially, the training labels are relying on the same or similar job titles which are not really adapted to the particular company. However, as the HR process continues, each new job description will be paired with a list of resumes shortlisted scored according to their reaching the first interview, second interview, with the actual candidate hired to have the highest score (1). This way we can collect a set of good labels for the model to be trained. As the fine-tuning of the BERT pretrained model is relatively fast, it is recommended, once we have collected sufficient samples for training, that older resumes / job descriptions be phased out from the training on a, say, quarterly basis, so that we can divest from older representations, say those created greater than 5 years ago, of the job definitions and keep the latest interpretation of fitness of a candidate’s resume.

# References

BIBLIOGRAPHY Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K., 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* [Online]   
Available at: https://arxiv.org/abs/1810.04805v2

Dhainje, A., 2019. *ResumeDataSet.* [Online]   
Available at: https://www.kaggle.com/dhainjeamita/resumedataset

Elmers, D., n.d. *The Job-Search Statistics All Job Seekers Should Know.* [Online]   
Available at: https://www.topresume.com/career-advice/7-top-job-search-statistics

Hab, M., 2017. *Online Job Postings - Dataset of 19,000 online job posts from 2004 to 2015.* [Online]   
Available at: https://www.kaggle.com/madhab/jobposts

Huggingface, n.d. *Transformers - State-of-the-art Natural Language Processing for PyTorch and TensorFlow 2.0.* [Online]   
Available at: https://github.com/huggingface/transformers

JobsPikr, 2019. *Australian job listings data from Gumtree - More than 8000 job listings extracted in 2018 from Gumtree.* [Online]   
Available at: https://www.kaggle.com/JobsPikrHQ/australian-job-listings-data-from-gumtree

Palan, M., 2019. *resumes.* [Online]   
Available at: https://www.kaggle.com/maitrip/resumes

PromptCloud, 2020. *Australian job listings data from Seek job board - 20k+ job listings extracted in 2018.* [Online]   
Available at: https://www.kaggle.com/PromptCloudHQ/australian-job-listings-data-from-seek-job-board

Rajapakse, T., 2020. *Simple Transformers.* [Online]   
Available at: https://github.com/ThilinaRajapakse/simpletransformers

Raman, V., 2019. *How I used NLP (Spacy) to screen Data Science Resume.* [Online]   
Available at: https://towardsdatascience.com/do-the-keywords-in-your-resume-aptly-represent-what-type-of-data-scientist-you-are-59134105ba0d

Siddhapura, A., 2020. *Resume Dataset - Annotated resume data to create text corpus for NER system.* [Online]   
Available at: https://www.kaggle.com/avanisiddhapura27/resume-dataset

Wu, O. T., 2019. *Resume Rater.* [Online]   
Available at: https://github.com/ongteckwu/Resume-Rater