



UE18CS334

Natural Language processing

Review 1

Title -Resume Screening

Arpit Nigam -PES2201800069
Jill Hansalia -PES2201800152
Manav Agarwal-PES22018000



Literature Survey -Paper 1

Title - End-to-End Resume Parsing and Finding Candidates for a Job Description using BERT

Authors- Vedant Bhatia, Prateek Rawat, Ajit Kumar

Published-Arxiv.org, 2018



Literature Survey -Paper 1

Scope-:

1. Exploring the feasibility of building a generic resume parser which performs across different formats
2. Building a heuristic and BERT based model to convert resumes to a standard format
3. Building an end-to-end resume parsing and data collection tool for employers in the form of a web application.
4. Ranking resumes as per their suitability to a job description using BERT sequence pair classification



Literature Survey -Paper 1

Methodologies-:

1. Building a Parser for Generic Resumes-:

For non standard cases, such as a two-column list format, the logical flow is not the same as the physical flow (left to right across the page). To tackle this issue of the logical flow of text, we built heuristic rules based on the spacing of text across the page. Larger spaces between words on the same horizontal line would indicate inter-column gaps whereas smaller spaces would indicate normal inter-word gaps. Clustering the extracted text into different headings

2. Building a Parser for a Standard Format of Resumes-:

For standard cases, the combination of the consistent structure and the metadata information available results in very successful heuristic based information extraction from these resumes.



Literature Survey -Paper 1

Methodologies-:

a. Converting Resumes to LinkedIn Format-:

To accomplish this task resumes were split into different segments based on the heuristic rules we had developed. We used BERT to create feature vectors for these segments. We train a classifier on the LinkedIn format resumes, and then classify each segment of the non-LinkedIn resumes into one of the LinkedIn format segment

3)Ranking candidates on the basis of job-description suitability-:

In order to simulate the job description that a company would be looking to hire individuals for, we used one candidate's description of his responsibilities at a previous role as the job description, positive samples by taking combinations of different job responsibilities that a person had and vice versa for negative samples.



Literature Survey -Paper 1

Advantages-:

- We used BERT for sentence pair classification and achieved 72.77% accuracy in predicting whether two job descriptions belong to same person or not.
- Takes both previous and next word's context, which couldn't be implemented before
- This method can be used to predict whether a person's previous job experience is similar to a job description at hand.
- This work establishes a strong baseline and a proof of concept which can lead to the hiring process benefiting from the advances in deep-learning and language representation.
- Through this method, we establish a strong baseline for candidate-job description suitability ranking



Literature Survey -Paper 1

Disadvantages-:

- Due to a lack of ground truth, we cannot train a neural network to rank resumes as per their suitability to the job description. Thus, use the sentence pair classification score from BERT as the ranking criterion, as this intuitively gives us a degree of similarity between the job description and profiles of candidates.
- A vision based page segmentation approach was missing in order to augment the structural understanding of resumes.
- Due to the bidirectional contextualisation, training of the model is extremely expensive and the model is limited to the English language



Literature Survey -Paper 2

Title -Automatic Extraction of Segments from Resumes using Machine Learning

Authors- Gunaseelan B, Supriya Mandal, Rajagopalan V

Published-IEEE, 2020



Literature Survey -Paper 2

Scope-:

- Extracting the skillset information from both the PDF and Word files.
- Extracting various features from text and building a classifier to predict whether a text line in a resume is heading or not which could be further used to build a classification model for the prediction of the category of each heading.
- Propose a system that uses multi-level classification techniques to automatically extract detailed segment information like skillset, experience and education from resume based on specific parameters



Literature Survey -Paper 2

Methodologies-:

- **Heading Prediction:** with the classifier algorithm. Multiple classification models have been built using various techniques, such as K-Nearest Neighbors (KNN), Support Vector Machine(SVM) with RBF kernel and ensembling techniques like Bagging,Boosting.
- **Segment Extraction:** All the extracted information is considered as the details of the previous heading. Then, we created a dictionary to store all section information in the key-value pair for every resume. Here key holds the title or heading and its respective content present in the value. This is because key should be unique as well as heading imparts
- **Segment Classification:** categorized 20 classes for all the major headings in resumes. We implemented the rule to extract the specific skill information from the dictionary, based on approximate string matching algorithm fuzzy-matching.

Literature Survey -Paper 2



DATA PREPARATION:-

- 1)Data Collection:- 400 Resumes Comprising of pdf and word files
- 2)Data Labelling:- Label 1 has been assigned to all the text, which is a heading and 0 for not-a-heading.
- 3)Data Preprocessing:- Excluding the table format resumes which didn't follow standard alignment.
- 4)Feature Extraction:- Word Count, Length of text, Text ends with symbol, Average word length, Stop-Words, Numeric, Special-Characters, Text-Case, POS, Similarity Features
- 5)Feature Analysis:- Count of label 0 is much more than that of label 1, & similarly the use of special characters is more frequent in class 1.
- 6)Feature Standardization:- The categorical features were transformed by applying one hot encoding technique, including bold and text ending with the special pattern.
- 7)Train/Test split:- Followed a 80-20 split



Literature Survey -Paper 2

Advantages-:

- We have successfully extracted the skillset from resumes, which gives 85% of accuracy from both the PDF and Word Files.
- This extracted skillset, experience and education can be used to recommend resumes for a job requirement.
- Manual annotation of data made the model to generalize well on unseen resumes.



Literature Survey -Paper 2

Disadvantages-:

- The dataset size was not large enough and only considered standard PDF and Word Files.
- XGBoost, with 5 cv fold, 1 hour to train on 1/10 of data, 6 to 12 hours to train on full dataset (depending n features used) thus the wait time was high



Literature Survey Paper-3

Title: Combination of Neural Networks and Conditional Random Fields for Efficient Resume Parsing

Author: Ayishathahira C H, Sreejith C, Raseek C

Published: IEEE-2018



Introduction

Resume parsing is a technique to extract useful information from resumes for further processing such as resume ranking and selection. This paper proposes a system for resume parsing using deep learning models such as the convolutional neural network (CNN), Bi-LSTM (Bidirectional Long Short- Term Memory) and Conditional Random Field (CRF).

CNN Model is used for classifying different segments in a resume. CRF and Bi-LSTM-CNN models were used for sequence labelling in order to tag different entities. Pre-trained Glove model is used for word embedding.



Architecture

Extracting Plain Text

Pre-processing

Segmentation: The segmentation model is created using a Convolutional Neural Network (CNN) architecture, which segments the information contained in the resume data into three different classes as personal, educational and occupational.

Information Extraction: Information extraction is the main task in resume parser, in which they extract useful information from each segment. A CRF based model and a Bi-LSTM-CNN model are created for sequence labelling section.



Results

- Segmentation model: 91%CNN, 43%Bi-LSTM
- Sequence labelling: CRF model gives accuracy of all classes in between 90%-95%.
Bi-LSTM-CNN model gives low accuracy of 65-70% for different classes.



Disadvantage

- Training time for CRF model is high for all classes due to high computational complexity.



Literature Survey Paper-4

Title: A System for Detecting Professional Skills from Resumes Written in Natural Language

Author: Emil St. Chifu, Viorica Rozina Chifu, Iulia Popa, Ioan Salomie

Published: IEEE-2017



Introduction

The proposed method uses an ontology of skills, the Wikipedia encyclopedia, and a set of standard multi word part-of-speech patterns in order to detect the professional skills.

First, the method checks to see if there are, in the text of the resumes, skills that are concepts in our ontology. The method also tries to identify possible new skills, which are not present in our ontology. This is done with the help of some specific, lexicalized, multi-word expression patterns.

The newly detected skills are validated by a human expert and then inserted automatically into the skill ontology.




Architecture

System consist of 4 modules: the crawler module, the specific pattern induction module, the skill detection module, and the ontology update module

The Crawler Module: This module visits specific Websites – sites that expose resumes to recruiters –, reads the pages, and extracts a large number of CV resumes in order to create a data corpus. This corpus will be used as training and testing data by our approach for detecting professional skills from text.

The Specific Pattern Induction Module: The module returns a set of lexicalized multi word expression patterns. These specific linguistic patterns will be used by the next module in the pipeline in order to detect new skills. There are two steps in the process of inducing the specific expression patterns: (i) text preprocessing and (ii) generating the specific multi-word expression patterns.

- 
- **The Skill Detection Module:** The main idea of this module is to identify and extract skills from a resume that are concepts in our skill ontology and to also detect possible skills that are not defined as concepts in the skill ontology. This process of noun phrase retrieval involves using the Stanford natural language processing framework. Every noun phrase is verified whether it contained as concept in ontology.
 - **The Ontology Update Module:** The goal of this module is to enrich the domain ontology of skills with newly added skills. New skill extracted from previous module is searched on Wikipedia for definition. Definition is splitted into subject and predicate. Then they are split into token and linking words are removed. Then these token are queried in skilled ontology and result is stored in map <concept,numberOfTimes>. Concept which appears more number of times then threshold is also added as they are possible taxonomic parents new skills. This is only done for predicate tokens.



Results

- The domain ontology of skills that is used by system consists of 13,337 concepts, which are split into two main categories: domain specific skills and competences and non-domain specific skills and competences, as well as relationships between them.
- The performance of the system is measured by using the most known information retrieval metrics: precision, recall, and Matthews correlation coefficient (The Matthews correlation coefficient (MCC) is a metric for classifying into two classes). These measures are computed based on the values of the true positives, true negatives, false positives, and false negatives.
- The values obtained for the Matthews correlation coefficient(0.6) show that the system offers a strong correlation between the expected output – in terms of correctly classifying a noun phrase as a skill versus as a non-skill.
- Average precision(0.58) and recall(0.62) is good but not perfect.



Disadvantage

Only 10 generic POS pattern is used in identifying new skills.

The sentences “I have experience with the C++ language” and “I have doubts about the future salary growth” both match the standard POS pattern “noun + preposition + article”. Yet the generic POS pattern will actually detect “the C++ language” as an unknown skill, (i.e. as a true positive) in the first sentence, whereas the same generic pattern will identify “I have doubts about the future salary growth” wrongly as a unknown skill(i.e. as a false positive).



Literature Survey Paper - 5

Title - NLP methods for automatic candidate's CV segmentation

Author - Maria Tikhonova, Anastasia Gavrishchuk

Published - IEEE - 2019

Ref - M. Tikhonova and A. Gavrishchuk, "NLP methods for automatic candidate's CV segmentation," 2019 International Conference on Engineering and Telecommunication (EnT), Dolgoprudny, Russia, 2019



Literature Survey Paper - 5

Introduction :

In this paper, The article proposes an algorithm for automatic CV parsing that is extraction information about work experience, education and basic info, which is based on Natural Language Processing methods. For each text line in a CV it is predicted whether it contains information about work experience or education and then CV segmentation into 3 blocks (basic Information, education and work experience) is performed. The testing dataset was taken from real CV's of candidates applying in Sberbank for different vacancies.



Literature Survey Paper - 5

Methodology :

The CV segmentation procedure consists of 7 consecutive steps:

Step 1. CV preprocessing and transformation - the algorithm runs through the text in a sliding window shifting one line down on each iteration and takes a text part.

Step 2. Word embeddings construction - On the second step words which are contained in CVs collection are embedded into the linear space R^n . Thus, for every word an n -dimensional vector is obtained.

Step 3. Computation of tf-idf index - tf-idf of a word t in document d which belongs to the collection of documents D is computed as the product of $tf(t, d)$ and $idf(t, d, D)$:



Literature Survey Paper - 5

Step 4. Text field embeddings construction - In this step, vector representations of the words present in the CV are summed with their tf-idf weights so as to give them an importance in the model.

Step 5. Specific features extraction - The feature space is expanded with specific features, which allow to increase classification results using 1) Parts-of-Speech counters 2) Counters of “Specific Suffixes”

Step 6. Text Line classification - The problem of CV segmentation is regarded as two binary classification tasks 1) Work Experience and 2) Education prediction.

Step 7. Final CV segmentation - All the above stated methods were applied to get a final classification of each line of the resume.



Literature Survey Paper - 5

Advantages :

On top of the classification results some rule-based heuristics were added:

- 1) Predictions are smoothed in a sliding window in order to get rid of outliers.
- 2) Small segments of the size less than 5 text windows are deleted.
- 3) When small segments are removed bounds of the remaining segments are adjusted.

Jaccard Index was used for evaluation of the classification.



Literature Survey Paper - 5

Disadvantages :

The training and testing Datasets taken for the model were taken from a website HeadHunter which provides resumes in a standardised format.

E.g. The Education part on HeadHunter contains only information about higher education without and courses or additional education. Thus, trained on such data classifier made errors classifying mention of courses in a CV, such as Coursera and more.



Literature Survey Paper - 5

Results :

The Jaccard Index was calculated for the test data set and in addition to this a CV was classified if it was well segmented or not as well based on the rule

$$J_{\text{work exp}}(C) + J_{\text{edu}}(C) > 1.7$$

$$J_{\text{card_work exp}} = 0.942$$

$$J_{\text{card_edu}} = 0.806$$



Literature Survey Paper - 6

Title - Automated Resume Evaluation System using NLP

Author - Rohini Nimbekar, Yogesh Patil, Rahul Prabhu, Shainila Mulla

Published - IEEE Xplore



Literature Survey Paper - 6

Introduction :

Resumes are unstructured documents based on the applicant's writing skills, they can be created in a multitude of formats. Dynamic extraction techniques are used to extract the most relevant information from the resumes. **So the companies need parsed resumes for achievement.** So if a company needs an associate employee who mainly incorporates a high experience level in python language, solely then his resume can be shortlisted.



Literature Survey Paper - 6

Scope :

A massive-scale company receives many resumes every day, thus handling those numbers of resumes has become a crucial task and time overwhelming method. Due to these reasons, numerous companies provide specific format for job seeker. Jobseeker ought to replenish with needed info then the CV / resume is going to be analyzed by machine. But this might become a problem for the applying candidates since they have to keep track of resume in different formats for every job they apply.

The focus in the selection process is on:

- (1) The method of selection and skills in terms of contribution to the dependability of selections.
- (2) The factor outlined and applied by decision-makers. And the way these reflect their comprehension of “necessary competence”.



Literature Survey Paper - 6

Methodology :

There are mainly 4 phases in the method undertaken in the paper:

A. Conversion Phase - The first process that comes into existence the conversion of unstructured data to structured data. Database consists of unstructured data, data in its negotiant state and successfully structured data.

B. Extraction Phase - Resumes have different structure and consist of many information like education, skills etc. Entity Name Extraction (NER) has been used in the transformation of unstructured text to structured text. is a subtask of the information extraction method. Once a CV is structured, then extracting the relevant information becomes a straightforward task



Literature Survey Paper - 6

C. Filtration - This process gives all the applicants who match the job description. To make the filtration process more efficient, a score is given to each resume to rank the applicant. Collaborative filtering is used to to predict the trend of selection.

D. Ranking - Ranking the resume can facilitate the organization to choose the candidate wisely and helps in better recruitment process in a shorter period. Our system uses a machine learning algorithm to build the applicant learning model and predict the recruiter's judgment once given the candidate's resume.

In the testing phase, the candidate application is applied to learn model for sorting, and then finalized rank candidates list is generated



Literature Survey Paper - 6

Advantages :

The system deals with unstructured resumes and hence is more useful.

The system consists of multiple modules so as to ease the process namely,

A)Section based segmentation - Used to extract candidates information

B)Filtration module - This refines list by removing the insignificant terms.

C)The third module takes a set of skills extracted from both resumes and job portals as input to classify them under their corresponding occupational categories.



Literature Survey Paper - 6

Disadvantages :

The rankings and weights for a segment of resume may differ from company to company.

The domain was restricted to engineering resumes.



Resume Screening

Aim-: Ranking Resumes based on Specific Job Skills

Dataset-: 2274 Resumes (Txt,Docx,Doc,PDF) Collected from
https://github.com/Msq-9/Extraction-of-Skills/tree/master/data/raw_resume



Approach

Library Requirements:- PYPDF2, python-docx, textract, nltk, pdfminer3, docx

- Extracting text from PDF and storing it in a text file.
- Removing Stop Words.
- Extracting Contact No. & email accounts
- Extracting Bi-gram words from the sentences
- Using Gensim word2Vec to get word embedding to compute similar words(in our project we acquire job skills from the given job)
- Calculating the score for a job using the count of skills and applying min-max scaling.
- Filtering the resumes based on top score for a particular job

Output (Screening Banking Resumes)

```
a=input("The parameter wanted:")  
(sk.sort_values(by=a, ascending=False)).head()
```

parameter wanted:bank

Resume_Id	bank	trade	teach	engineer	driver	writer	data	cloud	programming	administration	sales	
resume1021	1.000000	0.071429	0.142857	0.0	0.133333	0.0	0.020134	0.000000	0.0	0.0	0.500000	[pnarula71@
resume481	0.319149	0.000000	0.000000	0.0	0.066667	0.0	0.006711	0.000000	0.0	0.0	0.318182	[perline_ta
resume302	0.319149	0.178571	0.000000	0.0	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.454545	[hamed@westviewcapita
resume980	0.297872	0.000000	0.000000	0.0	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.045455	[Rebecca.Dengx
resume685	0.276596	0.071429	0.000000	0.3	0.066667	0.0	0.013423	0.266667	0.0	0.0	0.181818	[ray.ls.yeur

Output (Finding resumes most qualified for programming)

```
a=input("The parameter wanted:")  
(sk.sort_values(by=a, ascending=False)).head()
```

The parameter wanted:programming

	Resume_Id	bank	trade	teach	engineer	driver	writer	data	cloud	programming	administration	sales	Email	Phone
697	resume697	0.042553	0.000000	0.285714	1.0	0.000000	0.333333	1.000000	1.000000	1.000000	0.0	0.181818	[NDCG@5, NDCG@10.]	[]
23	resume23	0.021277	0.000000	0.000000	0.6	0.000000	0.000000	0.026846	0.000000	0.307692	0.0	0.000000	[xu_zhang@live.com]	[]
471	resume471	0.000000	0.000000	0.000000	0.5	0.000000	0.000000	0.100671	0.266667	0.230769	0.0	0.000000	[jajjanyani@gmail.com]	[9772881151]
1014	resume1014	0.021277	0.178571	0.000000	0.6	0.066667	0.000000	0.060403	0.133333	0.230769	0.0	0.181818	[ANTONIO.LAVOURA@GMAIL.COM]	[]
1006	resume1006	0.021277	0.000000	0.142857	0.1	0.000000	0.000000	0.026846	0.000000	0.230769	0.0	0.136364	[anniewongchingyin@gmail.com]	[]