Taylor & Francis
Taylor & Francis Group

Journal of Discrete Mathematical Sciences and Cryptography

ISSN: 0972-0529 (Print) 2169-0065 (Online) Journal homepage: https://www.tandfonline.com/loi/tdmc20

A novel firefly driven scheme for resume parsing and matching based on entity linking paradigm

Gerard Deepak, Varun Teja & A. Santhanavijayan

To cite this article: Gerard Deepak, Varun Teja & A. Santhanavijayan (2020) A novel firefly driven scheme for resume parsing and matching based on entity linking paradigm, Journal of Discrete Mathematical Sciences and Cryptography, 23:1, 157-165, DOI: 10.1080/09720529.2020.1721879

To link to this article: https://doi.org/10.1080/09720529.2020.1721879

Published online: 20 Apr 2020.

Journal of Discrete Mathematical Sciences & Cryptography
ISSN 0972-0529 (Print), ISSN 2169-0065 (Online)
Vol. 23 (2020), No. 1, pp. 157–165
DOI : 10.1080/09720529.2020.1721879

Taylor & Francis
Taylor & Francis Group

A novel firefly driven scheme for resume parsing and matching based on entity linking paradigm

Gerard Deepak *
Varun Teja †
A. Santhanavijayan §
Department of Computer Science and Engineering
National Institute of Technology Tiruchirappalli
Tiruchirappalli 620015
Tamil Nadu
India

Abstract

In this paper, contemporary Natural Language Processing techniques have been leveraged to demonstrate the capability of data-driven HR towards significant improvement in the quality and speed of the whole recruiting process. Firstly, by using NLP, a resume parser has been implemented to analyze the most crucial recruitment parameters. Thereafter, ability to display a pie chart for a candidate has been employed in the algorithmic structure of the parser to prepare a powerful tool for the resume matching based on job criteria. To determine the efficacy and accuracy of the proposed resume ranker, an enhanced rival modern optimizer, i.e., firefly ranking algorithm is applied to accelerate the speed of ranking algorithm. An overall accuracy of 94.19% has been achieved by the proposed approach. The results indicate that the resume parser has been incorporated with robust techniques and hence concedes to the accuracy of the results.

Subject Classification: *03B52 , 68T20 , 68T50*

Keywords: *Data-driven HR, Firefly algorithm, NLP, Resume parser, Resume matching*

*E-mail: gerard.deepak.christuni@gmail.com (Corresponding Author)
†E-mail: varunteja866@gmail.com
§E-mail: vijayana@nitt.edu

1. Introduction

In today's world of huge competition, thousands of resumes flood into reputed companies for a vacant position. Although resumes have structured format, their structure is not well defined like other documents such as letters and circulars. Traditional method of manually skimming and scanning through the resumes are error prone and is highly unsuitable for corporate environments where recruiting talented professionals highly influences the growth of the company. The extraction of texts from the document and the order and form of information in it is of primary concern in parsing resumes. To overcome these issues, a three-level hierarchical structure consisting of segments, blocks and chunks has been assumed. The named entities in these chunks are used in processing information.

Contribution: In contrast to several non- structured document types, the resumes are reasonably structured. In resumes information are stored in separate blocks. Thus, the semi-structured characteristics of the resume have been utilized by parsers based on NLP for classification and extraction of information from resumes. Experiments revealed that a high accuracy has been achieved in segmentation which is a vital phase in resume parsing. The identification rates of named entities are within the acceptable ranges. However, an overall accuracy of 94.19% has been achieved where experiments are mainly focused on Indian resumes.

The remaining paper is organized as follows. Section 2 provides a brief overview of related r work. Section 3 describes the System Architecture. Section 4 discusses the Implementation, Results, and Performance Evaluation. Finally, the paper is concluded in Section 5.

2. Related Work

Das et al., [1] have addressed the problem of Named Entity Extraction based on big data tools for parsing resume. However, Sanyal et al., [2] have also parsed resume based on a pure Natural Language Processing scheme which involved POS tagging with syntax tree generation and chunking. Berty et al., [3] have proposed segmentation based on headers specifically for parsing resume in Indonesian. Matthew et al., [4] incorporated entity linking based on Convolutional Neural Networks using semantic similarity. Bhagavatula et al., [5] have used Web Tables scheme for performing entity linking. Moreover Wang et al., [6] have proposed Entity linking based on quantified validation which is domain independent. Hua et al., [7] have imbibed social context for arriving at a strategic scheme for

entity linking. In [8-15] various ontology focused applications have been discussed.

3. Proposed System Architecture

The architecture of the Proposed Resume Matching system is depicted in Figure 1. There are 4 phases in the parser system namely, Text segmentation, Named Entity Recognition, text normalization Co-reference, Merging and Conflict resolution. Text Segmentation is the most crucial phase in parsing of resume. In this phase, the extracted text is split into segments of related information. The resume is split into several segment based on the attributes like Name, Phone, Email and Web information are clustered under the segment Contact. Each block in resume is given a heading, using a data dictionary that store common headings in a resume. Appropriate segments could be identified and stored in database by mapping them with the headings A group of named entity recognizers has been designed to work only for specific segments and are termed as chunkers. Hence particular group of chunkers work only for fixed segments. Once segmentation is done, the segmented outputs are subject to Named Entity Recognition.

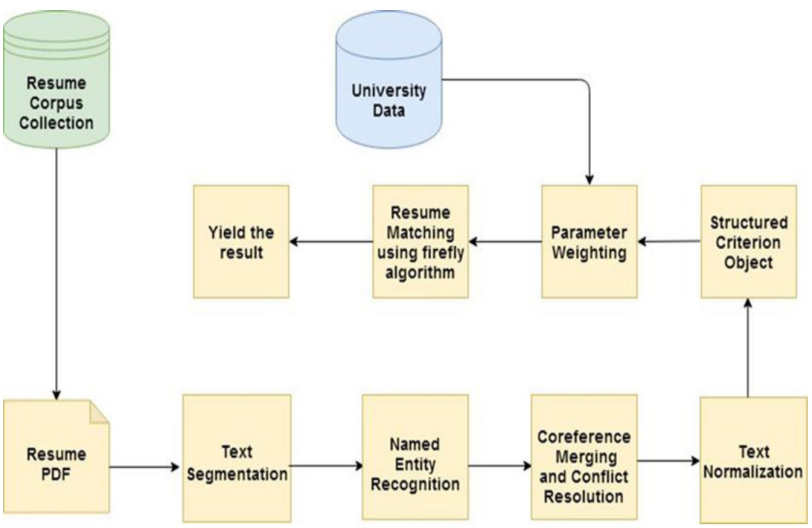

Figure 1

The architecture of the proposed resume matching and parsing system

There are exclusively designed chunkers to recognize each named entity. For example, "is-headquarter-of" is used between an organization and location and such properties are used by chunkers to recognize the named entities. The chunkers are designed using four types of information, namely, the cue words, eminent names, prefixes and suffixes of words and finally the manner of writing name of a person. The cue words include prepositions. The texts are then classified into named entities. The chunkers then produce the output in a form such that separate contact, work-information and skillsets are separately segregated. Furthermore, once Named Entity Recognition is complete, the parsed resume is subject to text normalization where the specific entities are transmuted to make them coherent and reliable. In this phase, abbreviations are expanded using a reference repository. The parsed resume is subject to Co-reference resolution which is a linguistic phenomenon whereby two or more linguistic expressions may represent or indicate the same entity. The contextual rules in a resume is straight-forward and hence, could be easily represented in a program. The co-reference resolution has been used to link proper names occurring in text to personal or company. In this stage, the output of the parser system has been analyzed and are used to fill table output by chunkers. Thus, variations on the same theme, should all map to the same data object. These data objects are built by searching the well-formed substring table after the parse is complete, and mapping the structures identified by the parser into the fields of the record. In the course of processing a document, the program may extract additional (possibly redundant) information about events that it has already encountered, and so such templates may need to be updated or merged. Merging the two data objects collates these two sources of information.

The parameter weighting criteria is one of the most important tasks that needs to be mandated. The weight is assigned to resumes based on their educational qualifications and skills. The sum total of the weights is used in matching the resumes of candidates. Further, the weighted, structured data from resume is subject to matching by classification based on firefly algorithm. Resumes extracted from the recruiter's database are given as input to the algorithm. Each resume is also assigned the weight based on the above parameters which are used to find set of resumes matching the job requirements. The crux of optimization is searching for the optimal solutions to a specific arithmetical or statistical problem of interest. This search activity can be accomplished using multiple techniques. Therefore, efficient optimization algorithm is used to search for a suitable candidate. Instead of solving the entire data set and finding global optimum, using

this algorithm decreases time complexity of searching the candidate with maximum weight. In this way, the process of ranking the resumes based on job requirements becomes faster.

4. Implementation, Results and Performance Evaluation

The algorithm has been successfully designed and implemented in Linux/Fedora operating system. The implementation was executed on Intel i5 processor and includes 8GB of RAM. The requests are handled using GitHub API web request. Re and NLTK are purposed for regex-based NLP tasks and for named entity clustering respectively. The entire implementation was established by utilizing the pool of resumes in the corpus collection to carry out the experimentation and analysis of the outcome. The following task was to tokenize the resumes via text segmentation phase. After carrying out this phase, the segments are caused to undergo named-entity recognition. The resumes that do not satisfy the minimum requirements are neglected. If there is at least one resume that satisfies the minimum requirement then the resumes are relatively ranked with this as reference. In subsequent iterations, the matching set of resumes has been found. The matching set consists of a pool of resumes of highly qualified candidates. The algorithm is clever incorporation of already existing firefly algorithm with resume matching and is indicated in Table 1.

The experiment was organized on dataset from the resume corpus collection. The dataset was garnered by a technique to incorporate web-based resume processing. The performance is gauged on the basis of the following metrics. The information retrieval systems are based on text classification that helps deciding the relevance of the extracted information under a specific category. The metric rates output by the implemented segmenter lies in the range of 0.95 and 0.99. This fact clearly indicates the accuracy of the segmenter. Such high accuracy is reached because the data repository used contains almost all commonly used headings in a resume. The accuracy % is a standard measure of efficiency of the NER. Table 2 indicates the accuracy % tested against each of the identifier entity. The named entity recognizer's accuracy is within the tolerable error limit. The error occurs mainly due to the varied formats of mentioning particulars such as University, email etc. In spite of these difficulties, the chunkers designed proved to be rationally efficient. The design of chunkers used for NER is founded on cue word, famous names, prefixes-suffixes and style of writing places, cities, names etc. The carefully devised chunker has proved

Table 1

Algorithm for Resume matching based on Firefly

Input: Heterogeneous Resume-Corpus collection, criteria for selection based on job requirement.

Output: The set of resumes that match the job requirements.

begin

Step 1: Resumes from recruiters database be denoted by R_i . Each criterion is denoted as C_i. R_i a two tuple. $R_i = (S_i, W_i)$ where R_i denotes the resume i, S_i is the unique id assigned to the resume, W_i is the weight assigned to resume in parameter weighting phase.

Step 2: Let the objective function be f(R), $R = (R_1, R_2,\ldots,R_d)$;

Step 3: Input the resumes from the recruiter's database R_i, i = 1, 2, …, n

Step 4: Formulate the weights W of the resume so that it is associated with f(R)

Step 5: Let there be a resume ranking coefficient γ which could be any constant. It is used to decrease the appropriateness of resumes as the proximity of the resumes to the criteria mentioned by organization decreases.

Step 6: t = 0, Max = No. of candidates to be selected;

```
while (t ≤ Max)
    for i = 1: n (All resumes)
    for j = 1: i (n resumes)
    if (W_i ≥ W_j),
        Vary appropriateness of resume with distance r via exp (-γr)
        Move resume i towards resume j
        Evaluate new solutions and update the set of resumes that
        match the criteria
        end if
    end for j
    Rank resumes and find the current best matching resume to
    the job criteria
end for i
t = t + 1;
Post processing the results and visualization
end while
end
```

its high accuracy. The text segmentation and NER are the most vital tasks in resume parsing. The high precision value of segmentation and NER has contributed significantly to the overall efficiency of the resume parser

Table 2
Performance Measures of Named Entity Recognition

Experimentation Entities	Recall %	Precision %	F-measure %	Accuracy %
Name	92.47	95.64	94.01	94.06
Qualification	92.81	95.77	94.32	94.3
Skillset	93.21	94.63	93.89	93.92
Mobile number	91.83	96.91	94.43	94.37
Email	91.47	93.82	92.69	92.645
Work Experience	93.33	95.87	94.57	94.6

6. Conclusions

The Resume parsing is proposed as an automated framework to demonstrate its relevance in corporate recruitment environments. The authenticity of the obtained results has been verified by conducting laborious comparative study on familiar data-sets. The experimental results indicated that the proposed Resume Parser can be used as a robust recruitment tool to cope easily with the resume screening difficulties and the following task of ranking the resumes of the candidates. An overall accuracy of 94.19% has been achieved by the proposed approach. The results of the current research give an inspiration to incorporate this parser in real-time applications. In future, a research work would be accomplished on the aspects of improvement of accuracy of segmentation phase in the proposed resume parser. As a scope of future work, the resume corpus collection could be enlarged and a research could be conducted on improving the accuracy of NER, thus enhancing the accuracy.

References

[1] Das, P., Pandey, M., & Rautaray, S. S. (2018). A CV Parser Model using Entity Extraction Process and Big Data Tools.

[2] Sanyal, S., Hazra, S., Adhikary, S., & Ghosh, N. (2017). Resume Parser with Natural Language Processing. *International Journal of Engineering Science,* 4484.

[3] Tobing, B.C.L., Suhendra, I.R. and Halim, C., 2019, June. Catapa Resume Parser: End to End Indonesian Resume Extraction. In

Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval (pp. 68-74). ACM.

[4] Francis-Landau, Matthew, Greg Durrett, and Dan Klein. "Capturing semantic similarity for entity linking with convolutional neural networks." arXiv preprint arXiv:1604.00734 (2016).

[5] Bhagavatula, Chandra Sekhar, Thanapon Noraset, and Doug Downey. "TabEL: entity linking in web tables." In *International Semantic Web Conference*, pp. 425-441. Springer, Cham, 2015.

[6] Wang, Han, Jin Guang Zheng, Xiaogang Ma, Peter Fox, and Heng Ji. "Language and domain independent entity linking with quantified collective validation." In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 695-704. 2015.

[7] `Hua, W., Zheng, K. and Zhou, X., 2015, May. Microblog entity linking with social temporal context. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data* (pp. 1761-1775). ACM.

[8] Deepak, Gerard, and J. Sheeba Priyadarshini. "Personalized and Enhanced Hybridized Semantic Algorithm for web image retrieval incorporating ontology classification, strategic query expansion, and content-based analysis." Computers & Electrical Engineering 72 (2018): 14-25.

[9] Deepak, Gerard, and Sheeba Priyadarshini. "A hybrid framework for social tag recommendation using context driven social information." *International Journal of Social Computing and Cyber-Physical Systems* 1, no. 4 (2016): 312-325.

[10] Deepak, Gerard, Ansaf Ahmed, and B. Skanda. "An intelligent inventive system for personalized webpage recommendation based on ontology semantics." *International Journal of Intelligent Systems Technologies and Applications* 18, no. 1-2 (2019): 115-132.

[11] Deepak, G., Shwetha, B.N., Pushpa, C.N., Thriveni, J. and Venugopal, K.R., 2018. A hybridized semantic trust-based framework for personalized web page recommendation. *International Journal of Computers and Applications*, pp.1-11.

[12] Pushpa, C. N., Gerard Deepak, J. Thriveni, and K. R. Venugopal. "Onto Collab: Strategic review oriented collaborative knowledge modeling using ontologies." In 2015 Seventh International Conference on Advanced Computing (ICoAC), pp. 1-7. IEEE, 2015.

[13] Giri, G.L., Deepak, G., Manjula, S.H. and Venugopal, K.R., 2018. OntoYield: A Semantic Approach for Context-Based Ontology Recommendation Based on Structure Preservation. In Proceedings of International Conference on Computational Intelligence and Data Engineering (pp. 265-275). Springer, Singapore.

[14] Deepak, G., and Z. Gulzar. "Ontoepds: Enhanced and personalized differential semantic algorithm incorporating ontology driven query enrichment." *Journal of Advanced Research in Dynamical and Control Systems* 9, no. Specia (2017): 567-582.

[15] Deepak, Gerard, and Dheera Kasaraneni. "OntoCommerce: an ontology focused semantic framework for personalised product recommendation for user targeted e-commerce." *International Journal of Computer Aided Engineering and Technology* 11, no. 4-5 (2019): 449-466.