# Resume Screening Using Natural Language Processing and Machine Learning: A Systematic Review


**Arvind Kumar Sinha, Md. Amir Khusru Akhtar, and Ashwani Kumar**

**Abstract** Curriculum vitae or resume screening is a time-consuming procedure. Natural language processing and machine learning have the capability to understand and parse the unstructured written language, and extract the desired information. The idea is to train the machine to analyze the written documents like a human being. This paper presents a systematic review on resume screening and enlightens the comparison of recognized works. Several techniques and approaches of machine learning for evaluating and analyzing the unstructured data have been discussed. Existing resume parsers use semantic search to understand the context of the language in order to find the reliable and comprehensive results. A review on the use of semantic search for context-based searching has been explained. In addition, this paper also shows the research challenges and future scope of resume parsing in terms of writing style, word choice and syntax of unstructured written language.

**Keywords** Machine learning · Natural language processing · Resume parser · Semantic search · Unstructured written language

## 1 Introduction

Resume parsing is the process that extract information from websites or unstructured documents using complex patterns matching/language analysis techniques. It is a means to automatically extract information from resumes/unstructured documents and to create a potential database for recruiters. This process generally converts free-form of resumes, that is, pdf, doc, docx, RTF, HTML into structured data such as XML or JSON. Artificial intelligence technology and natural language processing (NLP) engine are used to understand human language and automation. Resume parsers use

A. K. Sinha (✉) · Md. Amir Khusru Akhtar
Usha Martin University, Ranchi, India
e-mail: passionarvind@gmail.com

A. Kumar
Vardhaman College of Engineering, Hyderabad, India



D. Swain et al. (eds.), *Machine Learning and Information Processing*,
Advances in Intelligent Systems and Computing 1311,
https://doi.org/10.1007/978-981-33-4859-2_21

semantic search to parse data from available resumes and find suitable candidates. The process of extracting human language is a difficult because human language is infinitely varied and ambiguous. A human language is written and expressed in several ways; thus parsing tool need to capture all the ways of writing by using complex rules and statistical algorithms. Ambiguity comes when the same word can mean different in different contexts. For example, a four-digit number may be a part of telephone number, a street address, a year, a product number or version of a software application. Thus, the idea is to train the machine to analyze the context of written documents like a human being.

Recruitment agencies use resume parsing tools to automate the process and to save recruiters hours of work. Resume parser automatically separates the information into various fields based on the given criteria. The relevant information extracted by a resume parser includes personal information (such as name, address, email), experience details (such as start/end date, job title, company, location), education details (such as degree, university, year of passing, location), hobby (such as dancing, singing, swimming) and so on. There are numerous choices for resume parsers such as Sovren, Textkernel, Rchilli, BurningGlass, Tobu, JoinVision CVlizer, Daxtra, HireAbility, RapidParser and Trovix [1]. Most companies use applicant tracking system which bundles resume parser as one of the features. The first resume parsers were used in the late 1990s as a stand-alone packaged solution for HR operation [2].

This paper presents a systematic review on resume screening and enlightens the comparison of recognized works. Several techniques and approaches of machine learning for evaluating and analyzing the unstructured data have been discussed. Existing resume parsers use semantic search to understand the context of the language in order to find the reliable and comprehensive results. A review on the use of semantic search for context-based searching has been explained. In addition to that, this paper also shows the research challenges and future scope of resume parsing in terms of writing style, word choice and syntax of unstructured written language.

The rest of the paper is organized as follows. Section 2 discusses information extraction methods. Section 3 presents a systematic review on resume parsers and enlightens the comparison of recognized works. Section 4 discusses the use of semantic search for context-based searching. Section 5 presents the research challenges and future scope of resume parsing. Finally, Sect. 6 concludes the paper.

## 2 Information Extraction Methods

Information extraction is the method of extracting definite information from textual sources. The textual information is divided into sentences called sentence segmentation or sentence boundary detection [3]. The rule-based approach [4] for segmentation uses a list of punctuation symbols such as '.', '?', ';' but this approach fails when it encounters abbreviations like 'e.g.', 'etc.', 'n.d.' and so on. In order to classify punctuation marks carefully supervised machine learning technique was proposed. It uses decision tree to mark the sentence boundaries and classification of punctuation

symbols [5]. The supervised machine learning approach requires huge corpora for training and needs specific knowledge of abbreviations [6]. Kiss and Strunk proposed unsupervised machine learning approach that uses type-based classification. In this method a word is analyzed in the whole text and annotated in sentence boundary and abbreviation annotation [6].

After segmentation of sentence boundaries, the system divides the sentence into tokens called tokenization. Several tokenization approaches have been proposed in the literature such as rule-based and statistical approaches. A rule-based tokenizer approach uses a list of rules for classification of tokens such as Penn Tree Bank (PTB) tokenizer [7]. Statistical approach uses hidden Markov model (HMM) [8] to identify the word and sentence boundaries [9]. This method uses scanning and HMM boundary detector modules for tokenization.

In order to identify the meaning of the word part of speech (POS), taggings such as the Penn Treebank Tagset (PTT) [7], CLAWS 5 (C5) Tagset [10] can be used. An important task in information extraction is name entity recognition (NER), which identifies names of entities such as group, persons, places, currency, ages and times [11].

## 3 Resume Parsers

Natural language processing and machine learning have the capability to understand and parse the unstructured written language, and extract the desired information. Existing resume parsers use semantic search to understand the context of the language in order to find the reliable and comprehensive results. A resume parser converts the unstructured form of data into a structured form. Resume parser automatically separates the information into various fields based on given criteria. It separates the information into various fields based on the given criteria and parameters such as name, address, email, start/end date, job title, company, location, degree, university, year of passing, location, dancing, singing and swimming [2]. Several open-source and commercial resume parsers are available for information extraction.

### *3.1 Open-Source Resume Parser*

Open-source resume parsers are distributed with source code and these sources are available for modification. These open-source libraries parse free-form of resumes, that is, pdf, doc, docx, RTF, HTML into structured data such as XML or JSON. Meanwhile, the social media profile links these parsers and parse the public webpages and convert these data into structured JSON format, such as LinkedIn and Github. Table 1 shows the list of open-source resume parsers and its properties.

These open-source parsers are simple and easy to use except Deepak's parser and follows the same approach for cleaning and parsing. These parsers still contain

**Table 1** Open-source resume parsers

| Resume Parser | Focuses on | Programming language | Library used | Output file | Advantage | Disadvantage |
|---|---|---|---|---|---|---|
| Brendan Herger's (Herger, 2015/2020) | Extracts information from resume | Python | PDFMiner | CSV file | Simple and Language dependent | Information loss in terms of date and job description |
| Skcript Technologies' (Skcript/Cvscan, 2016/2020) | Extracts information from resume | Python | PDFMiner | .json format | Simple and CLI interface | Fails to extract date most of the time |
| Antony Deepak's (GitHub—Antonydeepak/resume parser: resume parser using rule based approach. developed using framework provided by GATE, n.d.) | Extracts information from resume uses a hybrid machine-learning and rule-based approach focuses on semantic rather than syntactic parsing | Java | Apache's Tika library | Structured.json format | Better accuracy | Complex and difficult to use |
| Keras-English-resume-parser-and-analyzer (Chen 2018/2020) | Extracts information from English resume use Keras and deep learning models | Python | PDF miner | Raw content | Simple and better accuracy | Language dependent |

**Table 2** Commercial resume parsers

| Resume parser | Focuses on | Output file | Advantage |
|---|---|---|---|
| HireAbility's ALEX [13] | Extracts information from resume | HR-XML, JSON | Supports multiple languages and locales<br>Accurate, fast and secure |
| RChilli's [14] | Extracts information from resume | XML, JSON | Self-learning capability<br>Fast and reliable |
| DaXtra [15] | Extracts information from resume | XML, JSON | Multilingual resume parsing<br>Most comprehensive and accurate |
| Rapidparser [16] | Extracts information from resume | XML, JSON | Multilingual resume parsing<br>simple and accurate |

HTML and Unicode characters with negative effect on named entity recognition [12].

### *3.2 Commercial Resume Parsers*

Commercial resume parsers are designed and developed for sale to end users [12]. These resume parsers have more classy algorithms for attribute recognition than open-source parsers and allows them to correctly identify these attributes. The strength of commercial parsers undoubtedly lies in the careful analysis of resume to identify different sections such as skill, qualification and experience sections. Table 2 shows the list of commercial resume parsers and its properties.

Many parsers are available in the market that provides CV automation solutions and round-the-clock customer support. These resume parsers APIs are inexpensive and easy to integrate.

## 4 Semantic Searches for Context-Based Searching

Semantic search means searching text with meaning to improve correctness of search by understanding the searcher intent [17]. In comparison to lexical search, the program looks for exact matches without understanding the meaning. Semantic search uses various parameters such as context, place, purpose and synonyms, to find appropriate search results. The benefits of semantic search with reference to resume screening include the ability to perform fuzzy matching, allow pattern recognition, fetch experience by context, capable to establish relation between words and

ideas. Context-based search includes various parts of search process such as understanding the query and knowledge. Literature shows the use of semantic search for context-based searching and are very effectives in parsing resume [18].

A semantic binary signature has been proposed in the literature [19]. It processes a search query by determining relevant categories and generates a binary hashing signature. The appropriate categories are examined and hamming distances are calculated between inventory binary hashing signatures and search query. The hamming distance shows semantic significance that can be used to understand the searcher intent.

A novel sentence-level emotion detection using semantic rules has been published in the literature [20]. This paper discusses an efficient emotion detection method and matches emotional words from its emotional_keyword database. This technique investigates the emotional words and provides better result and performance than existing researches.

An NLP-based keyword analysis method has been proposed in the literature [21]. This method uses three matrices document content matrix V, word feature matrix W and document feature matrix H. Then, rank is calculated for each word using the set of coefficients. Finally, rank is generated for one or more queries using the ranks for each word.

Thomas and Sangeetha proposed [22] an intelligent sense-enabled lexical search on text documents to extract word from text document. This method uses word sense disambiguation (WSD) of each word and then semantic search on the input text to extract semantically related words. This method of extraction is useful in resume screening, resume learning and document indexing.

Alexandra et al. proposed [23] design and implementation of a semantic-based system for automating the staffing process. The proposed system uses skills and competencies lexicon for semantic processing of the resumes and matches the candidate skills as per the job necessities. This method eliminates reputative activities to minimize processing time of recruiter and improves search efficiency using complex semantic criteria.

Kumar et al. [24–27] proposed an object detection method for blind people to locate objects from a scene. They used machine-learning-based methods along with single SSMD detector algorithm to develop the model.

This research shows the uses of semantic search in order to understand the context of the language for reliable and comprehensive results.

## 5 Research Challenges and Future Scope

The correctness of resume parser depends on a number of factors [1], such as writing style, choice of words and syntax of written text. A set of statistical algorithms and complex rules are needed to suitably know and fetch the correct information from resumes. Natural language processing and machine learning have the capability to

understand and parse the unstructured written language and context-based information. There are many ways to write the same information such as name, address and date. So, resume parsing is still in its natal stage, and few important challenges and future scope are as follows [1, 12]:

- Understanding writing style of resume
- Understanding choice of words in a resume
- Understanding syntax of unstructured written language
- Context-based searching
- Understanding organization and formatting of resume
- Understanding headers and footers of resume
- Breaking resume into sections
- Understanding the structural and visual information from PDFs
- Speed of parsing.

## 6 Conclusions

Resume screening is the process that extract information from unstructured documents using complex patterns matching/language analysis techniques. Natural language processing and machine learning have the capability to understand and parse the unstructured written language and context-based information. This paper presents a systematic review on resume screening and enlightens the comparison of recognized works and investigate open-source and commercial resume parser. This paper discusses several open-source and commercial resume parsers for information extraction. Then, a review on the use of semantic search for context-based searching has been explained. In addition, this paper also shows the research challenges and future scope of resume parsing in terms of writing style, word choice and syntax of unstructured written language.

## References

1. Résumé parsing, https://en.wikipedia.org/w/index.php?title=R%C3%A9sum%C3%A9_parsing&oldid=921328084 (2019)
2. Seiv, M., HR software companies? Why structuring your data is crucial for your business?, https://medium.riminder.net/hr-software-companies-why-structuring-your-data-is-crucial-for-your-business-f749ecf3255a. Accessed on 25 Jan 2020
3. Dale, R., Moisl, H., Somers, H., *Handbook of Natural Language Processing* (CRC Press, 2000)
4. Reynar, J.C., Ratnaparkhi, A., A maximum entropy approach to identifying sentence boundaries, in *Proceedings of the Fifth Conference on Applied Natural Language Processing* (Association for Computational Linguistics, 1997), pp. 16–19
5. Riley, M.D., Some applications of tree-based modelling to speech and language, in *Proceedings of the workshop on Speech and Natural Language* (Association for Computational Linguistics, 1989), pp. 339–352

6. T. Kiss, J. Strunk, Unsupervised multilingual sentence boundary detection. Comput. Linguist. **32**, 485–525 (2006)
7. The Stanford Natural Language Processing Group, https://nlp.stanford.edu/software/tokenizer.shtml. Accessed on 05 Mar 2020
8. Manning, C.D., Manning, C.D., Schütze, H., *Foundations of Statistical Natural Language Processing* (MIT Press, 1999)
9. B. Jurish, K.-M. Würzner, Word and sentence tokenization with hidden Markov models. JLCL **28**, 61–83 (2013)
10. UCREL CLAWS5 Tagset, https://ucrel.lancs.ac.uk/claws5tags.html. Accessed on 05 Mar 2020
11. L. Derczynski, D. Maynard, G. Rizzo, M. Van Erp, G. Gorrell, R. Troncy, J. Petrak, K. Bontcheva, Analysis of named entity recognition and linking for tweets. Inf. Process. Manag. **51**, 32–49 (2015)
12. Neumer, T., *Efficient Natural Language Processing for Automated Recruiting on the Example of a Software Engineering Talent-Pool* 88 (2018)
13. Resume parsing software | CV parsing software—HireAbility, https://www.hireability.com/. Accessed on 09 Mar 2020
14. Inc, Rc., Looking for a perfect job/resume parser alternative, https://www.rchilli.com/looking-for-a-perfect-job/resume-parser-alternative. Accessed on 09 Mar 2020
15. Resume Parsing Software | CV Parsing Software, https://www.daxtra.com/resume-database-software/resume-parsing-software/. Accessed on 09 Mar 2020
16. CV Parsing Lightning-fast - RapidParser, https://www.rapidparser.com/. Accessed on 09 Mar 2020
17. Semantic search, https://en.wikipedia.org/w/index.php?title=Semantic_search&oldid=940652635 (2020)
18. Bast, H., Buchhold, B., Haussmann, E., Semantic search on text and knowledge bases. Found. Trends® Inf. Retr. **10**, 119–271 (2016). https://doi.org/10.1561/1500000032
19. Liu, M.: Search system for providing search results using query understanding and semantic binary signatures, https://patents.google.com/patent/US20200089808A1/en (2020)
20. Seal, D., Roy, U.K., Basak, R., Sentence-level emotion detection from text based on semantic rules. In: Tuba, M., Akashe, S., Joshi, A. (eds.) *Information and Communication Technology for Sustainable Development* (Springer, Singapore, 2020), pp. 423–430. https://doi.org/10.1007/978-981-13-7166-0_42
21. Baughman, A.K., Diamanti, G.F., Marzorati, M., Natural language processing keyword analysis, https://patents.google.com/patent/US10614109B2/en (2020)
22. Thomas, A., Sangeetha, S., Intelligent Sense-enabled lexical search on text documents. In: Bi, Y., Bhatia, R., Kapoor, S. (eds.) *Intelligent Systems and Applications* (Springer International Publishing, Cham, 2020), pp. 405–415. https://doi.org/10.1007/978-3-030-29513-4_29
23. Alexandra, C., Valentin, S., Bogdan, M., Magdalena, A.: Leveraging lexicon-based semantic analysis to automate the recruitment process. In: Ao, S.-I., Gelman, L., Kim, H.K. (eds.) *Transactions on Engineering Technologies* (Springer, Singapore, 2019), pp. 189–201. https://doi.org/10.1007/978-981-13-0746-1_15
24. Kumar, A., A review on implementation of digital image watermarking techniques using LSB and DWT, in *Information and Communication Technology for Sustainable Development* (Springer, 2020), pp. 595–602
25. Kumar, A., Reddy, S.S.S., Kulkarni, V., An object detection technique for blind people in real-time using deep neural network, in *2019 Fifth International Conference on Image Information Processing (ICIIP)* (IEEE, 2019), pp. 292–297
26. A. Kumar, Design of secure image fusion technique using cloud for privacy-preserving and copyright protection. Int. J. Cloud Appl. Comput. IJCAC **9**, 22–36 (2019). https://doi.org/10.4018/IJCAC.2019070102
27. A. Kumar, S. Srivastava, Object detection system based on convolution neural networks using single shot multi-box detector. Proc. Comput. Sci. **171**, 2610–2617 (2020)