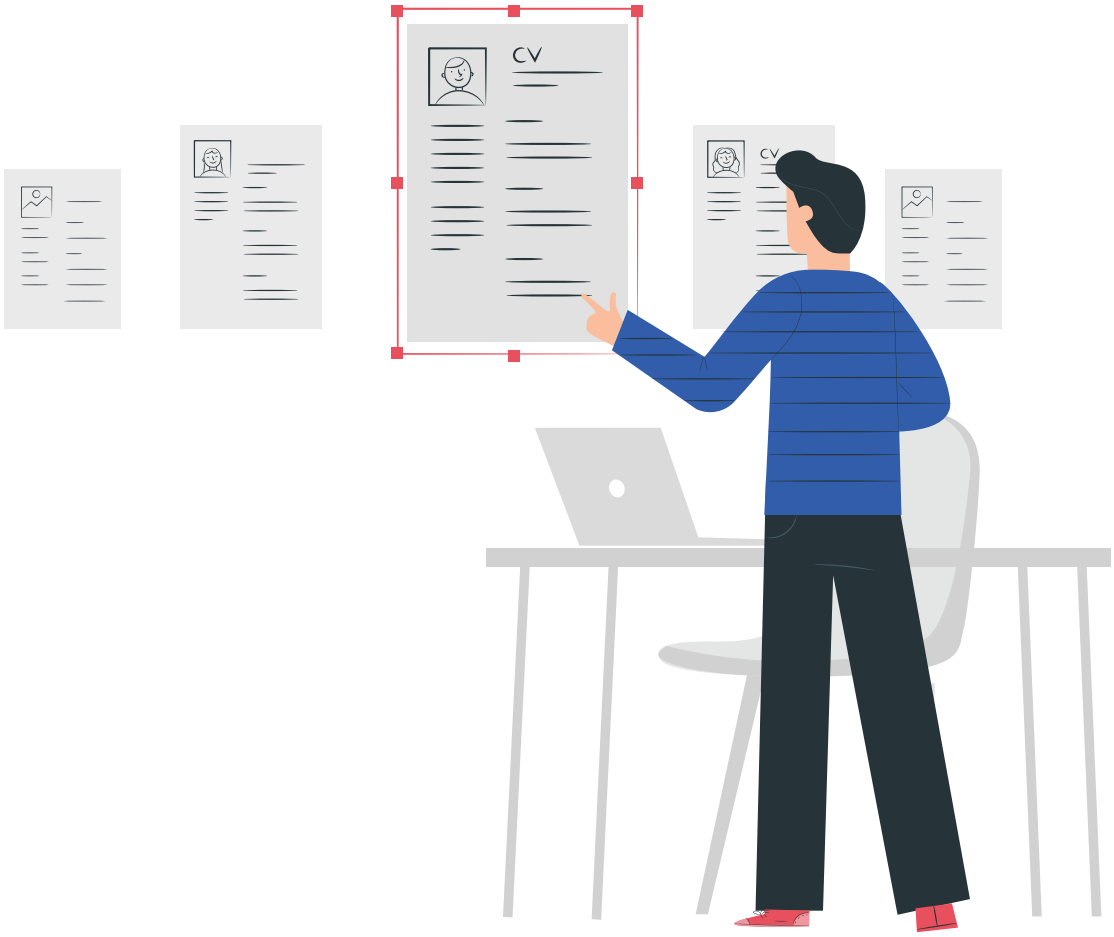
Skillate

AI Resume Parser: Fad or Fact?

An Insight into the working of ML-based Resume Parsing Technology



Recruiters have been screening resumes manually for a long time now. They read through every candidate resume and evaluate them on the basis of skills, knowledge, abilities and otherdesired factors.

However, it would take a long time for the recruiter to go through each resume in detail. So, in practical world recruiters are forced into doing one of the two things:

- They go through limited resumes, scan them thoroughly and take a pick out of them.
- They go through all (or most) of them, take a minimal amount of time to review them(some claim as low as **6 seconds**), and pick whichever resumes can hook them.

In both the cases, organizations lose out on quality candidates and recruiters waste their time and effort.

So, How can one avoid it?
This is where **resume parsers** come into the picture.

What is resume parsing?

Resume Parsing, formally speaking, is the conversion of a free-form CV/resume document into structured information - suitable for storage, reporting, and manipulation by a computer.

Resume parsers analyze a resume, extract the desired information, and insert the information into a database with a unique entry for each candidate. Once the resume has been analyzed, a recruiter can search the database for keywords and phrases and get a list of relevant candidates.

So why is resume parsing so difficult?

Almost everyone tries to use a unique template to put information on their CV, in order to stand out. For a human, reading these CVs or a job ad is an easy task. These semi-structured documents are usually separated into sections and have layouts that make it easy to quickly identify important information.

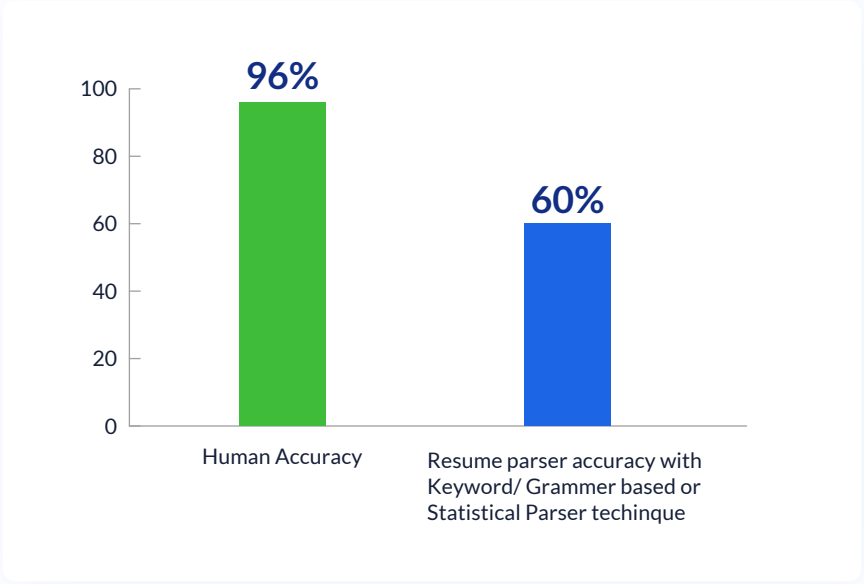
In contrast, for a computer, the task of extracting information becomes difficult with every change of format. Generally speaking here are the few kinds of resume parser available in the market:

01 **Key-word based resume parser:** A keyword-based resume parser works by identifying words, phrases, and simple patterns in the text of the CV/Resume and then applying simple heuristic algorithms to the text they find around these words

02 **Grammar based resume parser:** The Grammar-based resume parsers contain an enormous number of grammatical rules that seek to understand the context of every word in the CV/resume. These same grammars also combine words and phrases together to make complex structures that capture the meaning of every sentence in the resume.

03 **Statistical Parsers:** This type of parser attempts to apply numerical models of text to identify structure in a CV/Resume. Like grammar-based parsers, they can distinguish between different contexts of the same word or phrase and can also capture a wide variety of structures such as addresses, timelines, etc.

Without diving deep into the benefits and limitations of each of them, let us talk about the bottom line - **accuracy**. Most resume parsers use any of the above technologies to provide close to **60%** accurate results in real-world scenarios. However, when compared with **human accuracy of 96%**, they are definitely lagging behind.



There are many ways to write dates and numerous job titles and skills appear every month. Someone's name can be a company name (e.g. Harvey Nash) or even an IT skill (e.g. Cassandra). The only way a CV parser can deal with this is to "understand" the context in which words occur and the relationship between them.

That is why a rule-based parser will quickly run into two big limitations:

- The rules will get quite complex to account for exceptions and ambiguity
- The coverage will be limited.

So what's the solution?

ML-based Resume Parsing

The problem of Resume Parsing can be broken into two major subproblems - **Text Extraction**, and **Information Extraction**. A state of the art resume parser needs to solve both these problems with the highest possible accuracy.

01 Text Extraction

Almost everyone tries to use a unique template to put information on their CV. Even the templates that might seem indistinguishable to the human eye, are processed differently by the computer.

This creates the possibility of hundreds of thousands of templates in which resume are written worldwide. Not all templates are straightforward to read. For eg. One can find tables, graphics, columns in a resume, and every such entity needs to be read in a different manner. Therefore it is easy to conclude that rule-based parsers do not stand a chance and an intelligent algorithm is required to extract text in a meaningful manner from raw documents (pdf, doc, docx, etc).

Solution

Any parser that aims to become a state-of-the-art technology in its niche needs to explore several libraries- pdf, doc, docx, etc. However, a single type of algorithm is not good enough to extract all these document formats.

A new classification system that segregates the resumes into different types, based on their template, and tackles each type differently, is the way forward. Some of the types are straightforward, but most of them (like the ones that contain tables, partitions, etc) require higher-order intelligence from the software.

For such complex types, **Optical Character Recognition** (OCR) along with **Deep NLP** algorithms on top can help in extracting the required text. For every problem, there is a hard way and a smart way. OCR is a very generic problem which has been researched upon and solved by the biggest tech companies in the world. Most of the technology is open source as well! Therefore, rather than building a deep learning model from scratch for OCR and NLP, the smart way was to use the power of open source and deploy an off the shelf model for the task.

With the help of classification algorithms to segregate the resumes, some modern players have been able to amalgamate different technologies and obtain the best of all, to build highly accurate and fast text extraction methods.

02 Information Extraction

A typical resume can be considered as a collection of information related to - **Experience**, **Educational Background**, **Skills**, and **Personal Details of a person**. These details can be present in various ways, or not present at all.

Keeping up with the vocabulary used in resumes is a big challenge. A resume consists of company names, institutions, degrees, etc. which can be written in several ways. For eg. **Skillate:: Skillate.com** - Both these words refer to the same company but will be treated as different words by a machine. Moreover, every day new companies and institute names come up, and thus it is almost impossible to keep the software's vocabulary updated.

Consider the following two statements:

1. 'Currently working as a Data Scientist at <Amazon> Skillate'
And,
2. 'Worked in a project for the client Amazon '

In the first statement, "**Amazon**" will be tagged as a company as the statement is about working in the organization.

But the latter "Amazon" should be considered as a normal word and not as a company. It is evident that the same word can have different meanings, based on its usage.

Solution

The above challenges make it clear that statistical methods like Naive Bayes are bound to fail here, as they are severely handicapped by their vocabulary and fail to account for different meanings of words. So can this seemingly hard problem be cracked? Deep Learning can do all the hard work for us! This approach is called Deep Information Extraction.

A thorough analysis of the challenges posed makes it evident that the root of the problem here is understanding the context of a word.

Consider the following statement

‘2000–2008: Professor at Universitatea de Stat din Moldova’

It is quite likely that you wouldn’t have understood the meaning of all the words in the above statement, but even if you don’t understand the exact meaning of these words, you can probably guess that since “professor” is a job title, “Universitatea de Stat din Moldova” is most likely the name of an organization. Now consider one more example, with two statements:

‘2000–2008: Professor at IIT Kanpur’
‘B.Tech in Computer Science from IIT Kanpur’

Here, IIT Kanpur should be treated as an Employer Organisation in the former statement and as an Educational Institution in the later. We can differentiate between the two meanings of ‘IIT Kanpur’ here by observing the context. The first statement has ‘Professor’ which is a Job Title, indicating that IIT Kanpur be treated as a Professional Organisation. The second one has a degree and major mentioned, which point towards IIT Kanpur being tagged as an Educational Organisation.

Applying Deep Learning to solve Information Extraction greatly helped us to effectively model the context of every word in a resume.

To be specific, **Named Entity Recognition (NER)** is the algorithm that deep learning can be applied to, for information extraction in resumes.

"**NER** is a subtask of information extraction that seeks to locate and classify named entity mentions in unstructured text into predefined categories such as the person names, organizations, locations, etc, based on context."

Through the examples mentioned above, it should be clear that NER is a very domain-specific problem, and thus it is required to build a deep learning model from scratch.

Building the deep learning model

For building a model from scratch, the first step is to decide the model architecture. Research papers and other literature on NLP indicate using LSTMs (a type of Neural Network) in the model, as it takes into account the context of a word in a statement. Once the entire architecture is agreed upon, one needs to start working on curating a dataset for model training and evaluation. This step is the most cumbersome process and needs to be thought out from a very early stage.

One of the most important things to consider is the data on which the system is being trained. The data needs to be unlabelled and should not cause more ambiguity. Online tools that can help in collaborating the manual annotation efforts within the team are also of big help.

Small POCs on shorter datasets should be the early path to success. Once the results start to show up, data labeling and further training of the system can provide the desired results.

The task of data labeling is often considered trivial and lowly, however, it actually gives an insight into the performance of the model which is not possible with any research paper. Below is a snippet from the NER model results. It shows how the model is able to recognize and differentiate the different meanings of the phrase 'IIT Kanpur' in different contexts. Each word has a corresponding label.

```
I have completed my B.Tech from IIT Kanpur   I am working as an Assistant Professor at IIT Kanpur
OTH OTH OTH OTH S-DEG OTH B-INS I-INS        OTH OTH OTH OTH OTH B-TIT I-TIT OTH B-COM I-COM
```

TIT - Designation, COM - Professional Organization, INS - Educational Institute, DEG - Degree, OTH - Other

Benefits of AI-based resume parsing

Processes various file formats: The AI-based resume parser can process all popular file types including PDF, DOC, DOCX, ZIP, giving candidates the freedom to upload their resume in any format.

Deciphers complex resumes: The AI-based parser recognizes and extracts information from divergent formats. Example: Tabular templates, image scanning, etc.

Machine learning for better accuracy: Optical Character Recognition (OCR) and Deep NLP algorithms to extract text from resumes.

Lightning-fast processing: The AI-enabled parser takes 1-3 seconds to process the most complex of the resumes.

Resume Quality Score: Indexes resume based on their pedigree with AI-backed score, irrespective of the job profile.

About Skillate AI resume parser

Skillate is an advanced decision-making engine to **make hiring easy, fast, and transparent**. The product helps in optimizing the entire value chain of recruitment, beginning from **creating the job requisition, to resume matching, to candidate engagement**.

> The Skillate Resume Parser uses Deep Learning to extract information from the most complex resumes. The AI-enabled parser takes 1-3 seconds to process the most complex of the resumes. Deep Learning applied to Named Entity Recognition (NER) for 93% accurate information extraction.

Skillate

AI Recruitment platform

Skillate is an advanced decision-making engine to make hiring **easy**, **fast** and **transparent**.

India Office

#2751, Ground Floor, 31st Main Rd,
1st Sector, HSR Layout, Bengaluru,
Karnataka, 560102

+91 70223 08814

contact@skillate.com

US Office

1160 Battery Street East, Suites 100,
San Francisco, California, USA 94111

+1 415 918 6004

contact@skillate.com

software AG LARSEN & TOUBRO YES BANK

sopra steria SAINT-GOBAIN