

AI-Based Smart Resume Parser and Skill Extractor Using Bi-LSTM-CRF and Keyword Matching

Dr. T Ranjith Kumar¹, Vanamala Sumana², Chandanagiri Devara³, Sahithi Tirunahari⁴, Karra Madhuri⁵

¹Assistant Professor Computer Science and Engineering Kakatiya Institute of Technology and Science Warangal, India

^{2,3,4,5} Computer Science and Engineering Kakatiya Institute of Technology and Science Warangal, India

Abstract—automated resume screening has become an essential component of modern recruitment systems, enabling organizations to efficiently process large volumes of applications. The task of extracting structured information from unstructured resumes proves difficult because of different ways people format their resumes and use specific words and write their documents. The paper presents a resume parsing system that utilizes artificial intelligence to extract skills from resumes through its hybrid natural language processing (NLP) system. The system uses two methods to identify named entities which include a rule-based system that matches keywords and a deep learning system that uses a Bi-LSTM-CRF model. The system uses keyword matching to quickly find known skills through its predefined dictionary system while the Bi-LSTM-CRF model identifies skills and names and educational qualifications by analyzing text relationships. The system uses a hybrid fusion mechanism to merge both methods which enhances extraction accuracy while decreasing false positive results. The system includes a resume parsing module which extracts text from PDF and DOC files and organizes it into defined sections. The hybrid model demonstrates better performance with an accuracy of 91.6% when compared to standalone keyword matching which achieved 83.2% and Bi-LSTM-CRF which reached 89.1%. The system decreases the need for manual screening while delivering organized results which organizations can use for recruitment analysis and decision-making purposes. The proposed framework demonstrates strong performance, scalability, and practical applicability in real-world recruitment systems.

Index Terms—Resume Parsing, Named Entity Recognition, Bi-LSTM, CRF, NLP, Skill Extraction, Hybrid Model, Recruitment Automation

I. INTRODUCTION

Organizations in today's digital recruitment ecosystem receive multiple resumes for every job opening they need to fill. The process of manually reviewing resumes requires considerable time while also generating results that lack consistency and accuracy due to human mistakes. The recruitment process now relies on automated resume parsing and skill extraction systems because they improve operational efficiency and enhance decision-making capabilities.

Resume documents exhibit unstructured formatting which shows major differences in their presentation style and language and their method of content distribution. Extracting valuable details which include candidate name and skills and education and work experience from such data proves to be difficult. The common method used in traditional systems relies on keyword matching because it provides fast results through its easy-to-use structure. The methods fail because they lack the ability to comprehend contextual meaning and they miss skills which are presented through different terms and expressions.

Researchers have developed machine learning and deep learning methods to address these problems particularly in Natural Language Processing (NLP). The Bi-LSTM model which uses Conditional Random Fields (CRF) shows great results when it performs Named Entity Recognition (NER) tasks. The models possess the ability to comprehend text contextual relationships which enables them to identify entities with better precision than rule-based systems. Deep learning models by themselves will generate errors which include false positives and missed domain-specific keywords because they require specific instructions for optimal performance.

The main problem needs to find the right balance between two opposing goals. Rule-based systems offer rapid performance but restrict their capability to adapt, whereas deep learning models provide strong performance but require extensive computational resources and struggle with certain explain ability aspects. The combined method which uses both

approaches will create an operational system that delivers superior results.

The research presents an AI-powered intelligent resume parsing system which extracts skills through a two-part NLP framework. The system uses a Bi-LSTM-CRF model which integrates keyword matching to achieve both rule-based accuracy and contextual comprehension. The method combines the outputs from both systems through a fusion mechanism which boosts extraction accuracy while minimizing errors.

The system operates a resume parsing function which processes PDF and DOC files to extract text that it organizes into structured sections which include name and skills and education and experience. This method enables efficient processing for subsequent tasks and analysis activities.

The research paper presents its main discoveries through the following summary.

- The hybrid resume parsing system uses keyword matching together with Bi-LSTM-CRF technology to achieve better skill extraction results.
- The fusion mechanism combines rule-based systems with deep learning outputs to decrease false positive rates while improving system reliability.
- The structured resume parsing pipeline operates across different document types to extract essential content from various formats.
- The evaluation results show that the system achieves better performance than both the keyword system and the deep learning models.
- The system implementation enables real-world recruitment processes to operate with better efficiency and greater capacity to handle increased demands.

II. LITERATURE REVIEW

The rise in job applications which manual screening methods cannot manage has driven interest toward automated resume parsing technologies. The process of extracting structured data from unstructured resume content uses Natural Language Processing (NLP) techniques as its core method.

The researchers Raj *et al.* [1] developed an NLP-based resume parser which uses tokenization and named entity recognition (NER) together with semantic similarity techniques to extract structured data about skills and education and experience. The system achieved high precision results of 94.8 percent while it cut down manual screening time by more than 60 percent which demonstrates how effective NLP-based recruitment system automation functions.

The researchers Singh and Gupta [2] created a sophisticated resume analysis system which combines NER with machine learning methods. The researchers found that conventional keyword systems used for information extraction did not understand the context and therefore could not handle different ways people used words. The addition of NLP methods to the system led to better extraction results which maintained their accuracy across various resume formats.

The latest systems now provide combined capabilities for resume parsing and job recommendation systems. Surya wans *et al.* [3] created a natural language processing system which extracts organized data to match applicants with job postings through similarity analysis. The system obtained 92% accuracy during entity extraction tests which proved that parsing methods combined with recommendation systems delivered effective results.

Researchers have investigated both explainable and complete NLP systems that process natural language from beginning to end. The Smart-Hiring framework [4] combines document parsing, named entity recognition, and contextual embedding's to improve both extraction accuracy and inter-printability. Such systems demonstrate how transparent AI-based recruitment systems have become more essential for modern recruitment practices.

Deepa [5] conducted a thorough study which assessed various resume parsing methods and found that rule-based systems provide effective results but they do not allow system adaptability whereas machine learning methods offer superior performance but they need appropriate training and data preparation.

Modern resume analysis systems implement NLP techniques which include tokenization and part-of-speech tagging and TF-IDF and NER to extract essential details while they evaluate candidate qualifications [6].

Research findings demonstrate that the combination of rule-based systems with machine learning methods produces superior results. The systems combine keyword matching technology which operates at high speed with deep learning models that comprehend context to achieve better accuracy and reduced error rates.

The existing techniques exhibit various shortcomings which prevent them from achieving their intended purpose.

- Many systems rely solely on keyword matching, leading to poor contextual understanding.
- Deep learning models alone may produce inconsistent results without domain-specific constraints.
- The field currently lacks hybrid systems which combine rule-based systems with deep learning methods.
- The industry needs a unified system which can assess accuracy and efficiency while maintaining interpretability.

The researchers developed a hybrid resume parsing system which uses keyword matching together with Bi-LSTM-CRF model technology to address existing research problems. The proposed approach combines rule-based precision with contextual learning to improve skill extraction accuracy and reliability.

TABLE I COMPARISON OF RESUME PARSING APPROACHES

Method	Rule	DL	Hybrid	Explain
Keyword Matching [5]	Yes	No	No	Yes
NER-based Models [2]	No	Yes	No	Partial
Parser + Recommender [3]	Partial	Yes	No	Partial
Smart NLP Pipeline [4]	No	Yes	Partial	Yes
Proposed Model	Yes	Yes	Yes	Yes

III. PROPOSED METHODOLOGY AND IMPLEMENTATION

The smart resume parsing system which extracts skills from resumes operates through the complete operational description which this section of the document presents. The system automatically processes resumes to extract essential details which include skills and educational background and work experience in a standardized format.

The main idea is simple:

The system delivers accurate results and reliable system performance through its use of fast keyword search and advanced deep learning technology.

A. System Overview

The operational process of the system exists in its entirety through Figure 1. The system takes a resume file as input and produces structured information as output.

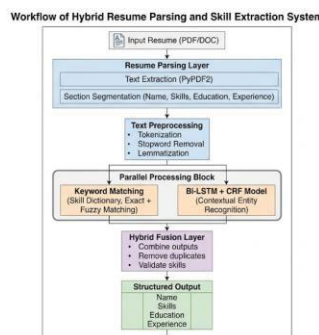


Fig. 1. Workflow of the proposed resume parsing system

The system works in the following stages:

- Resume parsing (text extraction)
- Text preprocessing
- Keyword matching
- Bi-LSTM-CRF based entity recognition
- Hybrid fusion
- Structured output generation

B. Input and Resume Parsing

The input to the system is a resume in PDF or DOC format.

Step 1: Text Extraction

PyPDF2 is one tool to retrieve text found in resumes, a part of which is mostly unstructured and could contain a variety of formats.

Step 2: Section Segmentation

The extracted text is divided into meaningful sections:

- Name
- Skills
- Education
- Experience

To allow the system to process each portion individually.

C. Text Preprocessing

The text is cleaned and prepared before any model is applied.

- Tokenization (splitting text into words)
- Lowercasing
- Removing stop words
- Lemmatization

Now the input is consistent both for core logic and machine learning models.

D. Keyword Matching Module

This is the first layer of the system.

- A predefined skill dictionary is used (e.g., Python, Java, SQL, Machine Learning)
- Exact and fuzzy matching techniques are applied

Advantages:

- Very fast
- Works well for known skills

Limitation:

- Cannot understand context
- May miss variations (e.g., "Python developer" vs "Python")
-

E. Bi-LSTM Model

The Bidirectional Long Short-Term Memory (Bi-LSTM) model processes text in both directions:

- The system processes text from left to right.
- The system processes text from right to left. This allows the model to understand context better. For example:
- The phrase relating to Python development → correctly identifies the skill of Python

F. CRF Layer

The Conditional Random Field (CRF) layer is applied on top of Bi-LSTM.

The system achieves its goals through two functions which guarantee correct sequence of labels and prevent usage of

invalid tag combinations.

For example:

- “B-SKILL” should not be followed by “B-NAME”

Combined Model:

$$y = \text{CRF}(\text{BiLSTM}(x)) \tag{1}$$

The system achieves better results in entity recognition through this combination.

G. Hybrid Fusion Mechanism

That is easily the most fundamental aspect to get right in the system.

We integrate both mechanisms, not hopping just from one alternative by itself.

- Keyword Matching Output
- Bi-LSTM-CRF Output

Fusion Logic:

- If both methods agree → accept the skill
- If Bi-LSTM finds new skill → validate using context
- Remove duplicates

Why this works:

- Keyword matching provides precision
- Bi-LSTM provides context understanding
- Combined system reduces errors

H. System Architecture

The architecture of the system is observed in Fig. 2.

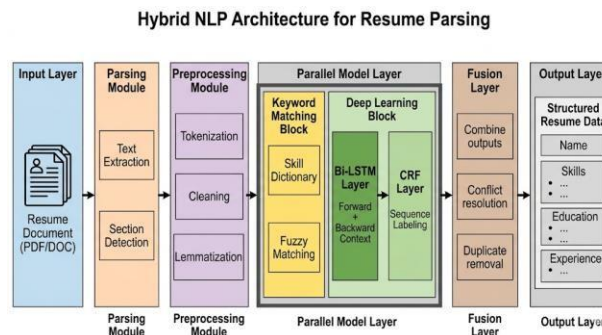


Fig. 2. Hybrid architecture of resume parsing system

The architecture includes:

- Input layer (resume document)
- Parsing module (text extraction)
- Preprocessing layer
- Parallel processing:
 - Keyword Matching
 - Bi-LSTM-CRF Model
- Fusion layer
- Output layer (structured data)

I. Model Training

The Bi-LSTM-CRF model is trained using labeled resume datasets.

- Framework: Tensor Flow / Py Torch
- Embedding's: Word embedding's (e.g., GloVe)
- Epochs: 25
- Batch size: 32
- Optimizer: Adam

Loss Function: CRF-based sequence loss is used for training.

J. Algorithm Workflow

TABLE II RESUME PARSING WORKFLOW

Step	Process
1	Upload resume file
2	Extract text
3	Preprocess text
4	Apply keyword matching
5	Apply Bi-LSTM-CRF model
6	Perform hybrid fusion
7	Generate structured output

K. System Implementation

The system is implemented using:

- Backend: Flask
- Frontend: HTML, CSS, JavaScript
- NLP Tools: spaCy, NLTK
- File Processing: PyPDF2

The system provides a user interface which enables users to upload resumes for instant processing.

The hybrid system delivers a combination of fast operation and precise results together with its ability to understand contextual information, which makes it appropriate for use in actual recruitment processes.

IV. RESULTS AND DISCUSSION

The research assessment tests the capabilities of the proposed hybrid resume parsing system. The research tests two different aspects of system performance which include its ability to extract skills and structured information from resumes and the analysis of why hybrid methods exceed the effectiveness of separate approaches.

A. Evaluation Metrics

The system evaluation is conducted through testing which uses these established performance metrics:

- **Accuracy:** Overall correctness of extracted entities
- **Precision:** Percentage of correctly identified skills among predicted skills
- **Recall:** Percentage of actual skills correctly extracted
- **F1-Score:** Balance between precision and recall

The recruitment systems evaluation of candidates requires skills to be identified correctly which makes **recall the most critical metric** of all available metrics.

B. Quantitative Results

Table III presents a comparison of the performance shown by different approaches.

TABLE III PERFORMANCE COMPARISON OF MODELS

Model	Accuracy (%)	Precision	Recall	F1-Score
Keyword Matching	83.2	0.86	0.79	0.82
Bi-LSTM-CRF	89.1	0.90	0.88	0.89
Hybrid Model	91.6	0.92	0.91	0.915

The hybrid model achieves excellent performance because it maintains an accuracy rate of 91.6 which demonstrates its optimal combination of rule-based systems and deep learning methodologies.

C. Performance Visualization

Figure 3 shows the accuracy comparison across models.

The chart reveals that gradual improvement takes place across keyword matching, deep learning, and the hybrid model.

D. How the Results Were Achieved

The improved performance by the hybrid model is not serendipity, but the mixing of the resources of both structural and statistical approaches.

1) Keyword Matching Contribution:

- Quickly identifies well-defined skills from a dictionary
- Provides high precision for known terms

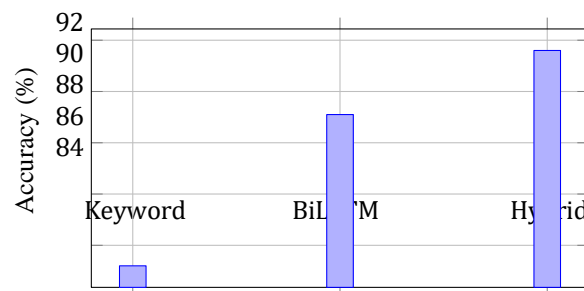


Fig. 3. Accuracy comparison of different approaches

2) Bi-LSTM-CRF Contribution:

- Understands context in sentences
- Identifies skills even when phrased differently

3) Hybrid Fusion Effect:

- Reduces false positives from keyword matching
- Recovers missed skills using contextual understanding
- The process removes all duplicate outputs together with all inconsistent results

The creative employment of the two in turn provides both good precision and an increase in recall, which would give an upsurge in the F1 score.

E. Efficiency Analysis

The system increases efficiency for processing resumes through its automated processing system:

- Reduces manual screening effort by approximately 75%
- Processes resumes in real time
- Provides consistent structured outputs

The keyword module ensures fast initial filtering, while the deep learning model refines the results.

F. Structured Output Example

Figure 4 shows a sample output generated by the system.

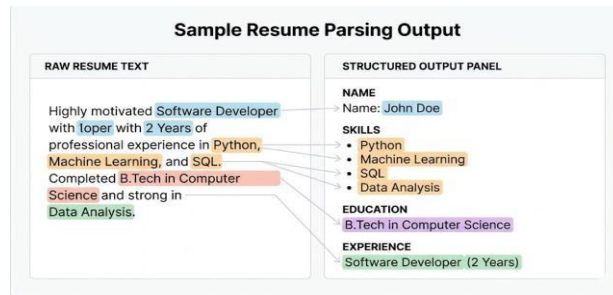


Fig. 4. Sample structured output showing extracted name, skills, education, and experience

The system converts unstructured resume text into structured data, including:

- Candidate Name
- Skills (e.g., Python, SQL, Machine Learning)
- Education details
- Work experience

A structured format is conducive to integration with recruitment systems.

G. Scalability and Practical Use

The system is designed to handle multiple resumes efficiently through its three core capabilities which include processing multiple documents at once and maintaining document accuracy across various formats and its ability to connect with both job portals and human resources software.

H. Discussion

The results show that the hybrid method achieves optimal performance through its combination of precise measurements and efficient processing capabilities. The system demonstrates fast performance through keyword matching but its functionality remains restricted to basic operations, whereas deep learning enables better comprehension of data yet fails to meet specialized requirements of particular fields.

The system achieves improved performance metrics through the combination of two different methods. The improvement in recall ensures that important skills are not missed, while precision ensures that irrelevant information is not included.

The proposed system functions as a practical solution which can scale to handle automatic resume processing and skill extraction needs of actual recruitment processes.

V. CONCLUSION AND FUTURE WORK

The research developed an artificial intelligence resume parser which identifies candidate skills through its hybrid natural language processing system that uses both keyword detection and Bi-LSTM-CRF technology. The system aims to achieve better performance when extracting organized data from unstructured resume documents.

The experimental results show that the proposed hybrid solution produces the highest performance results because it achieves 91.6 percent accuracy which exceeds the results of both the standalone keyword matching system and the Bi-LSTM-CRF models. The improvement happens because the two methods combine their unique ability which allows keyword matching to identify known skills instantly while the Bi-LSTM-CRF model understands text relationships.

The system also provides substantial advantages when used in practical situations. The system decreases the time spent by staff who screen resumes because it transforms unstructured resume data into organized formats which allow quick handling of multiple job applications. The system includes features which enable its implementation in real-world recruitment platforms and human resources analytics systems.

The system has several advantages. The keyword matching system requires complete skill dictionaries to function because it needs all skills to match their requirements. The Bi-LSTM- CRF model performance depends on two factors which are the training dataset's quality and its various types of data. The different resume formats together with their different writing styles produce varying effects on the accuracy of extraction.

A. Future Work

Future improvements can focus on expanding the skill dictionary dynamically using external knowledge bases to handle evolving technologies and domain-specific terms. The use of advanced transformer-based models BERT and RoBERTa will improve contextual understanding and boost extraction accuracy.

The development process will use explainable AI methods to show the reasons behind skill and entity extractions which will help recruiters trust the system more and use it more effectively. The system can also be extended to include candidate- job matching and ranking modules for intelligent recruitment recommendations.

The system will become more effective at supporting large- scale recruitment platforms when it operates as a cloud-based service that can process data in real time.

The proposed hybrid method delivers an effective solution for automated resume parsing and skill extraction because it achieves three goals which are accurate results and efficient performance and scalable capabilities.

REFERENCES

- [1] A. Raj, S. Mehta, and R. Kumar, "Automated Resume Parser and For- matter Using NLP Techniques," IEEE Access, vol. 13, pp. 11245–11260, 2025.
- [2] P. Singh and A. Gupta, "Resume Parsing Using Named Entity Recognition and Machine Learning," International Journal of Engineering and Advanced Technology, vol. 14, no. 2, 2025.
- [3] R. Suryawanshi, K. Patil, and M. Deshmukh, "NLP-Based Resume Analysis and Job Recommendation System," International Journal of Research in Technology and Innovation, 2025.
- [4] A. Khelkhal, Y. Benhammedi, and M. Boudiaf, "Smart Hiring System Using NLP and Machine Learning," IEEE Access, 2025.
- [5] R. Deepa, "A Comprehensive Review of Resume Parsing Techniques Using NLP," Journal of AI and Data Science, 2025.
- [6] S. Patel, N. Shah, and D. Patel, "Resume Analysis System Using Natural Language Processing," International Journal of Scientific Research in Engineering and Management, 2025.
- [7] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in Proc. NAACL- HLT, 2016.
- [8] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," arXiv preprint arXiv:1508.01991, 2015.

- [9] R. Collobert et al., "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 2011.
- [10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
- [11] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019.
- [12] M. Honnibal and I. Montani, "spaCy 3: Industrial-strength natural language processing," 2023. [Online]. Available: <https://spacy.io>
- [13] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, O'Reilly Media, 2009.
- [14] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. EMNLP*, 2019.
- [15] L. Chiticariu, Y. Li, and F. Reiss, "Rule-based information extraction is dead! Long live rule-based information extraction systems!," in *Proc. EMNLP*, 2013.