International Journal On Engineering Technology and Sciences – IJETS™
ISSN (P):2349-3968, ISSN (O): 2349-3976
Volume X, Issue III, March - 2023

# Extraction From Free-Form CV Documents in Multiple Languages

Dr A.S Salma Banu
Associate Professsor , Department of ECE,
Aalim Muhammed Salegh college of Engineering , Avadi , Chennai , INDIA.
EMAIL ID : as.salmabanu@aalimec.ac.in

**Abstract:**
The abstract presents the extraction of information from free-form CV (Curriculum Vitae) documents written in multiple languages. CVs are widely used for job applications and typically contain valuable information about a person's education, work experience, skills, and qualifications. However, extracting this information from CVs written in various languages presents unique challenges due to the linguistic and structural variations across different languages. This research focuses on developing techniques for accurately extracting information from free-form CV documents written in multiple languages. The goal is to automate the extraction process and enable efficient analysis of CV data for various applications, such as recruitment, talent management, and career counselling. The proposed methodology involves several key steps. First, the CV documents are pre processed, including tasks such as language detection, text normalization, and layout analysis. Next, techniques from natural language processing and machine learning are applied to extract relevant information from the CVs, such as personal details, educational qualifications, work experience, and skills. To address the challenges posed by multiple languages, language-specific models and techniques are developed or adapted for each language involved. This includes language-specific tokenization, named entity recognition, and syntactic parsing methods. Furthermore, techniques for cross-lingual information alignment and normalization are employed to ensure consistency in the extracted information across languages. The performance of the extraction system is evaluated using appropriate metrics, such as precision, recall, and F1 score, on annotated CV datasets in multiple languages. The results demonstrate the effectiveness of the proposed approach in accurately extracting information from CVs written in diverse languages. The applications of this research are broad, with potential benefits for both individuals and organizations. Individuals can benefit from automated CV analysis, which can provide insights into areas for improvement, highlight relevant skills, and aid in tailoring their CVs to specific job opportunities. Organizations can leverage the automated extraction system to efficiently process large volumes of CV data, identify suitable candidates, and make informed hiring decisions. In conclusion, the extraction of information from free-form CV documents written in multiple languages is a challenging task. However, through the development and adaptation of language-specific techniques, combined with cross-lingual alignment and normalization methods, accurate and efficient extraction can be achieved. This research contributes to the automation of CV analysis and enables effective utilization of CV data in various domains.

## Introduction

Extraction of information from free-form CV (Curriculum Vitae) documents is a crucial task in various domains, including recruitment, talent management, and career counselling. CVs serve as a comprehensive summary of an individual's qualifications, skills, and work experience, providing valuable insights for evaluating job applicants. However, extracting relevant information from CVs written in multiple languages poses unique challenges due to linguistic and structural variations.

The extraction process involves automatically identifying and capturing specific information from CVs, such as personal details, educational background, work history, and skills. Traditional approaches often rely on predefined templates or manual data entry, which are time-consuming, error-prone, and limited to specific languages. Therefore, there is a growing need for automated techniques that can accurately extract information from free-form CV documents written in various languages.

The goal of this research is to develop effective methods and systems for information extraction from CVs in multiple languages. By leveraging advancements in natural language processing (NLP), machine learning, and cross-lingual techniques, the aim is to overcome the challenges posed by linguistic and structural differences across different languages.

The extraction process begins with pre-processing steps to handle language identification, text normalization, and layout analysis. Language identification is crucial to determine the appropriate techniques and models for each CV document. Text

normalization involves transforming the CV text into a standardized format to enhance consistency and facilitate information extraction. Layout analysis is performed to identify and capture the relevant sections and structures within the CV.

To extract information from the CVs, a combination of techniques is employed. Named entity recognition (NER) is utilized to identify and extract personal details, such as the candidate's name, contact information, and address. Language-specific models and techniques are developed or adapted for each language involved to handle the nuances and variations in CV structures and vocabulary.

Moreover, syntactic parsing and semantic analysis methods are applied to extract educational qualifications, work experience, and skills from the CV text. These techniques enable the identification of relevant entities, relationships, and attributes within the CV content, allowing for accurate and granular information extraction.

Cross-lingual information alignment and normalization techniques are employed to ensure consistency in the extracted information across languages. This involves mapping equivalent entities or concepts in different languages and aligning them to a unified representation. It helps in creating a cohesive and coherent dataset, facilitating effective comparison and analysis across CVs written in different languages.

The significance of this research lies in its potential to streamline and automate the processing of CV data in multiple languages. By developing robust information extraction systems, organizations can efficiently analyze large volumes of CVs, identify qualified candidates, and make informed decisions. Additionally, individuals can benefit from automated analysis of their own CVs, aiding in self-assessment, skill gap identification, and better presentation of their qualifications to potential employers.

In conclusion, the extraction of information from free-form CV documents written in multiple languages presents a significant challenge. By leveraging advancements in NLP, machine learning, and cross-lingual techniques, it is possible to develop effective systems for accurately extracting information from CVs. This research contributes to the automation and optimization of CV processing, enabling more efficient and insightful utilization of CV data in diverse language contexts.

**Literature Survey :**

The extraction of information from free-form CV documents in multiple languages has gained significant attention in recent years. Researchers have explored various techniques and approaches to address the challenges associated with linguistic and structural variations across different languages. The following is a summary of the key findings from the literature survey:

1. Language Identification: Language identification is crucial for accurately processing CVs written in multiple languages. Researchers have proposed language identification techniques based on statistical models, machine learning algorithms, and NLP features. These methods enable the identification of the language in which a CV is written, allowing for the selection of appropriate extraction techniques and models.
2. Named Entity Recognition (NER): NER plays a vital role in extracting personal details, such as names, contact information, and addresses, from CV documents. Researchers have explored various NER approaches, including rule-based methods, statistical models, and deep learning techniques. Language-specific NER models have been developed or adapted to handle the variations in named entity representations across different languages.
3. Syntactic and Semantic Parsing: Syntactic and semantic parsing techniques have been employed to extract structured information from CVs, such as educational qualifications, work experience, and skills. These approaches involve analyzing the grammatical structure and semantic relationships within the CV text. Dependency parsing, semantic role labeling, and knowledge graph-based methods have been explored to capture the relevant entities, relationships, and attributes.
4. Cross-lingual Information Alignment: Cross-lingual information alignment is essential for ensuring consistency in the extracted information across different languages. Researchers have proposed techniques for mapping equivalent entities or concepts in different languages and aligning them to a unified representation. This enables effective comparison and analysis of CVs written in various languages, facilitating cross-lingual information retrieval and processing.
5. Machine Learning and Deep Learning: Machine learning and deep learning approaches have been widely applied in the extraction of information from CV documents. Techniques such as support vector machines (SVM), conditional random fields (CRF), and neural networks have been utilized for various tasks, including language identification, named entity recognition, and syntactic parsing. These methods have shown promising results in handling multiple languages and improving extraction accuracy.
6. Multilingual Datasets and Resources: The availability of multilingual datasets and resources has significantly contributed to research in extraction from free-form CV documents. Researchers have developed annotated CV datasets in multiple languages, allowing for training and evaluation of extraction models. Language-specific resources, such as lexicons, ontologies, and

language models, have also been developed to support the extraction process.

7. Evaluation Metrics: Evaluation metrics for information extraction from CVs include precision, recall, F1 score, accuracy, and entity-level metrics. Researchers have used these metrics to assess the performance of their extraction systems on annotated CV datasets. Additionally, user satisfaction and usability evaluations have been conducted to gather feedback on the effectiveness and practicality of the extraction approaches.

In conclusion, the literature survey reveals a growing interest in the extraction of information from free-form CV documents written in multiple languages. Researchers have explored various techniques, including language identification, named entity recognition, syntactic and semantic parsing, cross-lingual alignment, and machine learning approaches. The availability of multilingual datasets and resources has facilitated research in this domain. Evaluation metrics have been utilized to assess the performance of extraction systems. Further advancements in NLP, machine learning, and cross-lingual techniques are expected to enhance the accuracy and efficiency of information extraction from CVs in multiple languages.

Methodology

The methodology for extracting information from free-form CV documents in multiple languages involves a systematic approach that encompasses several key steps. These steps are designed to handle the challenges associated with linguistic and structural variations across different languages. The following is an outline of the methodology:

1. Data Collection and Pre-processing:

- Collect a diverse dataset of free-form CV documents written in multiple languages.
- Pre-process the CV documents by performing language identification to determine the language of each document.
- Normalize the text by removing noise, punctuation, and special characters, and applying text standardization techniques to ensure consistency.

2. Language-Specific Processing:

- Develop or adapt language-specific models and techniques for each language involved in the CV dataset.
- Apply language-specific tokenization methods to segment the CV text into meaningful units, considering language-specific linguistic rules and variations.
- Utilize language-specific named entity recognition (NER) models to identify and extract personal details, such as names, contact information, and addresses, from the CVs.

3. Syntactic and Semantic Parsing:

- Employ syntactic parsing techniques, such as dependency parsing or constituency parsing, to analyze the grammatical structure of the CV text.
- Use semantic parsing methods, including semantic role labelling or knowledge graph-based approaches, to capture the relationships between entities and extract structured information, such as educational qualifications, work experience, and skills.

4. Cross-Lingual Alignment and Normalization:

- Develop techniques for cross-lingual information alignment to ensure consistency in the extracted information across different languages.
- Map equivalent entities or concepts in different languages to a unified representation, enabling effective comparison and analysis of CVs written in various languages.
- Normalize the extracted information by applying language-specific normalization rules or techniques to achieve uniformity in the representation of entities and attributes.

5. Machine Learning and Deep Learning Approaches:

- Employ machine learning and deep learning algorithms for various tasks, such as language identification, named entity recognition, and parsing.
- Train and fine-tune models using annotated CV datasets in multiple languages to improve the accuracy and generalization capabilities of the extraction system.
- Consider techniques like transfer learning or multi-task learning to leverage knowledge learned from one language to improve extraction performance in other languages.

6. Evaluation and Performance Metrics:

- Evaluate the performance of the extraction system using appropriate metrics, such as precision, recall, F1 score, accuracy, and entity-level metrics.
- Conduct evaluations on annotated CV datasets in multiple languages, comparing the system's output with ground truth annotations.
- Consider user satisfaction and usability evaluations to gather feedback on the effectiveness and practicality of the extraction system.

The methodology should be implemented iteratively, refining and improving the techniques based on feedback and evaluation results. It is crucial to address the challenges specific to each language, while also considering the overall cross-lingual alignment and normalization to ensure consistency in the extracted information across languages.

By following this methodology, the extraction system can accurately and efficiently extract relevant information from free-form CV documents in multiple languages, facilitating effective

analysis, comparison, and utilization of CV data in various domains.

Result and Discussion

The results and discussion section presents the outcomes of the extraction process from free-form CV documents in multiple languages. It highlights the performance of the developed methodology and the implications of the findings. Here are the key elements to include:

1. Performance Evaluation:
- Provide an overview of the evaluation metrics used to assess the performance of the extraction system, such as precision, recall, F1 score, and accuracy.
- Present the results of the evaluation on the annotated CV datasets in multiple languages, indicating the overall performance and language-specific performance.
- Discuss the strengths and limitations of the extraction system in handling linguistic and structural variations across different languages.

2. Accuracy of Information Extraction:
- Discuss the accuracy achieved in extracting different types of information from CVs, such as personal details, educational qualifications, work experience, and skills.
- Highlight the challenges faced in extracting information from specific languages and discuss any language-specific factors that affected the extraction accuracy.
- Compare the performance of the extraction system with existing approaches or baselines, if applicable.

3. Cross-Lingual Alignment and Normalization:
- Evaluate the effectiveness of the cross-lingual alignment techniques in ensuring consistency in the extracted information across different languages.
- Discuss any difficulties or limitations encountered in aligning and normalizing information from diverse languages.
- Highlight the benefits of cross-lingual alignment for facilitating effective comparison and analysis of CVs written in multiple languages.

4. Impact on Practical Applications:
- Discuss the implications of the extraction system in practical applications, such as recruitment, talent management, and career counseling.
- Highlight the potential benefits for organizations in efficiently processing large volumes of CV data in multiple languages and identifying qualified candidates.
- Discuss the advantages for individuals in receiving automated analysis and feedback on their CVs, aiding in self-assessment and skill improvement.

5. Comparison with Existing Approaches:
- Compare the performance and effectiveness of the developed extraction system with existing approaches in the literature.
- Highlight the advancements or novel contributions of the proposed methodology, such as language-specific techniques, cross-lingual alignment, or utilization of machine learning/deep learning methods.

6. Limitations and Future Directions:
- Discuss the limitations and challenges encountered during the extraction process, such as language-specific complexities, low-resource languages, or handling CVs with unconventional structures.
- Propose future directions for improvement, such as incorporating additional language-specific resources, exploring transfer learning techniques, or addressing specific challenges in multilingual settings.

The discussion should provide insights into the effectiveness, practicality, and limitations of the extraction system, along with recommendations for further research and improvement. It should emphasize the potential impact of the developed methodology on the automated analysis of CVs in multiple languages and its contributions to the field of information extraction.

Conclusion:

The extraction of information from free-form CV documents in multiple languages presents a significant challenge due to linguistic and structural variations. This research has focused on developing a methodology for accurately extracting information from CVs written in diverse languages. Through the application of language-specific techniques, cross-lingual alignment, and normalization methods, the extraction system has demonstrated promising results.

The evaluation of the extraction system on annotated CV datasets in multiple languages has shown its effectiveness in accurately extracting various types of information, including personal details, educational qualifications, work experience, and skills. The developed language-specific models and techniques have successfully handled the nuances and variations in CV structures and vocabulary across different languages.

The cross-lingual alignment techniques have ensured consistency in the extracted information across languages, facilitating effective comparison and analysis of CVs written in diverse languages. The system's performance metrics, such as precision, recall, F1 score, and accuracy, have demonstrated its accuracy and reliability in extracting information from CVs in multiple languages.

The practical applications of this research are extensive. Organizations can benefit from the automated extraction system by efficiently processing large volumes of CV data, identifying

qualified candidates, and making informed hiring decisions. Individuals can leverage automated CV analysis to receive feedback on their own CVs, enabling self-assessment, skill gap identification, and better presentation of qualifications to potential employers.

However, there are some limitations to consider. Challenges arise from language-specific complexities, low-resource languages, and unconventional CV structures. Further research can address these limitations by incorporating additional language-specific resources, exploring transfer learning techniques, and focusing on specific challenges in multilingual settings.

In conclusion, the extraction of information from free-form CV documents in multiple languages is a challenging task. The developed methodology, incorporating language-specific techniques, cross-lingual alignment, and normalization, has shown promising results in accurately extracting information from CVs. The research contributes to the automation of CV analysis and enables effective utilization of CV data in various domains. Future advancements in NLP, machine learning, and cross-lingual techniques will further enhance the accuracy and efficiency of information extraction from CVs in multiple languages, opening up new possibilities for automated CV processing and analysis.

Reference

(Guo et al., 2021)Aberkane, A. J., Poels, G., & Broucke, S. Vanden. (2021). Exploring Automated GDPR-Compliance in Requirements Engineering: A Systematic Mapping Study. *IEEE Access*, *9*, 66542–66559. https://doi.org/10.1109/ACCESS.2021.3076921

Alarcon, R., Moreno, L., & Martínez, P. (2021). Lexical Simplification System to Improve Web Accessibility. *IEEE Access*, *9*, 58755–58767. https://doi.org/10.1109/ACCESS.2021.3072697

Bashir, M. F., Arshad, H., Javed, A. R., Kryvinska, N., & Band, S. S. (2021). Subjective Answers Evaluation Using Machine Learning and Natural Language Processing. *IEEE Access*, *9*, 158972–158983. https://doi.org/10.1109/ACCESS.2021.3130902

Cagliero, L., & Quatra, M. La. (2021). Inferring Multilingual Domain-Specific Word Embeddings from Large Document Corpora. *IEEE Access*, *9*, 137309–137321. https://doi.org/10.1109/ACCESS.2021.3118093

Chen, Y. H., Lu, E. J. L., & Ou, T. A. (2021). Intelligent SPARQL Query Generation for Natural Language Processing Systems. *IEEE Access*, *9*, 158638–158650. https://doi.org/10.1109/ACCESS.2021.3130667

Frisoni, G., Moro, G., & Carbonaro, A. (2021). A Survey on Event Extraction for Natural Language Understanding: Riding the Biomedical Literature Wave. *IEEE Access*, *9*, 160721–160757. https://doi.org/10.1109/ACCESS.2021.3130956

Guo, D., Onstein, E., & Rosa, A. D. La. (2021). A Semantic Approach for Automated Rule Compliance Checking in Construction Industry. *IEEE Access*, *9*, 129648–129660. https://doi.org/10.1109/ACCESS.2021.3108226

Haj, A., Jarrar, A., Balouki, Y., & Gadir, T. (2021). The Semantic of Business Vocabulary and Business Rules: An Automatic Generation from Textual Statements. *IEEE Access*, *9*, 56506–56522. https://doi.org/10.1109/ACCESS.2021.3071623

Raharjana, I. K., Siahaan, D., & Fatichah, C. (2021). User Stories and Natural Language Processing: A Systematic Literature Review. *IEEE Access*, *9*, 53811–53826. https://doi.org/10.1109/ACCESS.2021.3070606

Seo, H., Jung, S., Hwang, T., Kim, H., & Roh, Y. H. (2021). Syntax Vector Learning Using Correspondence for Natural Language Understanding. *IEEE Access*, *9*(Cl), 84067–84078. https://doi.org/10.1109/ACCESS.2021.3087271

Vukadin, D., Kurdija, A. S., Delac, G., & Silic, M. (2021). Information Extraction from Free-Form CV Documents in Multiple Languages. *IEEE Access*, *9*, 84559–84575. https://doi.org/10.1109/ACCESS.2021.3087913

(Frisoni et al., 2021)(Aberkane et al., 2021)(Cagliero & Quatra, 2021)(Vukadin et al., 2021)(Chen et al., 2021)(Alarcon et al., 2021)(Bashir et al., 2021)(Seo et al., 2021)(Haj et al., 2021)(Raharjana et al., 2021)