IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Entity Recognition By Natural Language Processing And Machine Learning

Ms. Sharmila Sunil Dalave,Prof. S. A. Gaikwad,

1Student, 2Assistant Professor

1Department of Computer Science and Engineering, TPCT'S COLLEGE OF ENGINEERING, OSMANABAD, India

*Abstract:*Named Entity Recognition (NER) plays a crucial role in Natural Language Processing (NLP) by identifying key entities such as persons, organizations, locations, and numerical values in text. This research presents a multilingual NER and translation system integrating SpaCy for entity extraction, Google Translate API for multilingual conversion, and Web Speech API for voice input processing. The system follows a Flask-based backend and React.js frontend, ensuring scalability and real-time processing. This paper discusses the architecture, implementation, challenges, and performance evaluation of the system.

Key Words: Named Entity Recognition (NER), Information Extraction, Information Retrieval.

INTRODUCTION

With the rise of global communication, businesses and researchers require efficient text processing systems capable of extracting meaningful information and translating it into multiple languages. Named Entity Recognition (NER) automates entity extraction, while machine translation ensures cross-lingual accessibility. This paper proposes an NER-based text processing system that integrates NLP techniques, real-time translation, and speech recognition for enhanced user interaction. The system leverages SpaCy's NLP models, Google Translate API, and Web Speech API to provide accurate, automated, and scalable multilingual text processing. It aims to improve efficiency in industries such as customer support, healthcare, and legal analysis by reducing manual effort and enhancing communication.

I. RELATED WORK

Several research studies have focused on improving Named Entity Recognition (NER) accuracy using rule-based, statistical, and deep learning approaches. Early rule-based methods relied on predefined linguistic patterns and dictionaries, which lacked flexibility and struggled with unseen words. To address this, statistical approacheslikeHidden Markov Models (HMMs) and Conditional Random Fields (CRFs) were introduced, improving generalization by leveraging probabilistic models. However, these methods still required extensive feature engineering and manual annotation, limiting scalability.

With the rise of deep learning**,** Bidirectional Long Short-Term Memory (BiLSTM) networks combined with CRFs became popular for sequence labelling tasks, significantly improving NER performance by capturing contextual relationships. More recently, Transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers), RoBERTa, and T5, have revolutionized the field by leveraging self-attention mechanisms to understand complex sentence structures and contextual dependencies. These models, trained on massive corpora, have achieved state-of-the-art performance on benchmark NER datasets like CoNLL-2003 and OntoNotes**.**

Additionally, machine translation systemshave evolved fromrule-basedandstatistical machine translation (SMT)toNeural Machine Translation (NMT). Early translation methods relied on handcrafted linguistic rules, but they struggled with complex grammar and idiomatic expressions. SMT models, such as IBM Model 1-5, introduced probabilistic phrase-based translations but often resulted in word-by-word translations without proper contextual meaning. NMT models, particularly Transforme
based architectures like Google's Transformer model, have demonstrated superior performance in context preservation and fluency. These models employ attention mechanisms, allowing them to focus on relevant parts of a sentence during translation.

II. PROPOSED METHODOLOGY

The proposed system follows a modular architecture that includes a React.js frontend, a Flask-based backend, SpaCy for NER processing, Google TranslateAPI for multilingual support, and Web Speech API for speech-to-text conversion. The workflow involves preprocessing input text, entity extraction, optional translation, and result visualization. A database can be incorporated for entity storage and retrieval.

3.1Various approaches to solving NER issues

Early Named Entity Recognition (NER) methods primarily relied on rule-based approaches, where entities were first defined and then extracted. With advancements in machine learning, modern NER techniques have evolved into supervised, semi-supervised, and unsupervised learning approaches. Supervised learning requires large-scale annotated datasets and includes models like Hidden Markov Models (HMM), Maximum Entropy (ME), and Conditional Random Fields (CRF) for sequence labeling. Semi-supervised learning improves performance by training on small annotated datasets combined with unlabeled data, using techniques such as self-training and distant supervision. Unsupervised learning, on the other hand, relies on clustering methods and lexical similarity tools like WordNet to classify entities. A practical NER system can be developed using linguistic grammar rules, statistical machine learning models, or a hybrid of both. This NER project utilizes SpaCy's pre-trained NLP models, ensuring efficient entity recognition in real time. Additionally, deep learning techniques such as Bidirectional LSTMs and Transformer-based models like BERT enhance accuracy by capturing contextual word relationships. The integration of Google Translate API provides multilingual translation, while the Web Speech API enables speech-to-text conversion, making the system adaptable for various real-world applications.

3.2Challenges in Named Entity Recognition

Despite being a fundamental component of Natural Language Processing (NLP), Named Entity Recognition (NER) faces several challenges due to the complexity and variability of human language. Some of the key challenges encountered in this NER project are as follows:

- **Ambiguity and Abbreviations:** One of the major challenges in NER is identifying entities that have multiple meanings depending on context. Words can be used differently across sentences, leading to confusion in classification. Additionally, abbreviations and acronyms further complicate entity recognition, as the same abbreviation may represent different entities in different domains.
- **Spelling Variations:** Variability in spelling due to regional differences, typos, or phonetic similarities makes it difficult for NER models to consistently identify entities. Some words may have minor spelling changes that significantly alter their meaning, requiring robust text normalization techniques to ensure accurate entity extraction.
- **Foreign Words and Multilingual Processing:** Recognizing named entities across multiple languages is another significant challenge, especially when dealing with transliterations, code-switching, or loanwords. Some names and locations might appear in different forms across languages, making multilingual entity recognition difficult. This project addresses this issue by integrating Google Translate API to enhance language adaptability.

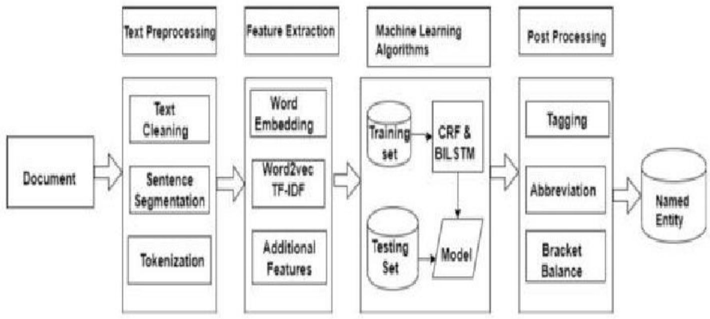

Fig. 1: Methodology

3.3Data Cleaning

The input data for Named Entity Recognition (NER) in this project comes from various sources such as documents, news articles, websites, social media, and Wikipedia, often in multiple languages. Since this data is usually unstructured, data preprocessing is essential for accurate extraction and classification of named entities. In this project, the SpaCy NLP library is used for Named Entity Recognition, enabling efficient preprocessing and entity extraction.

To obtain structured data, preprocessing techniques are applied, including sentence segmentation, tokenization, lemmatization, stopword removal, and part-of-speech tagging. The following steps ensure that the text is properly cleaned before being processed by the NER model:

- **Tokenization:** The process of breaking text into sentences and words to enable further analysis.
- **Sentence Segmentation:** Divides paragraphs into sentences, making it easier for the NER model to analyze entities contextually.
- **Lemmatization:** Converts words to their base form, ensuring consistency in entity recognition (e.g., "running" → "run").
- **Stemming:** Reduces words to their common root form, helping to normalize variations of the same word.

By applying these data-cleaning techniques, the project ensures that the NER model efficiently identifies and classifies entities, leading to more accurate recognition and translation of named entities across multiple languages.

3.4 Chunking and POS Tagging

Chunking and Part-of-Speech (POS) tagging play a crucial role in Named Entity Recognition (NER) by structuring unstructured text and improving entity extraction. In this project, SpaCy's NLP pipeline is used for POS tagging and chunking to enhance NER accuracy.

- **Chunking:** Chunking is the process of extracting meaningful phrases from text using POS tags as input. It helps identify structured segments such as noun phrases (NP) and verb phrases (VP), which are essential for extracting named entities like locations, personal names, and organizations. In this project, chunking is applied to recognize entity groups, improving entity classification and contextual understanding.
- **POS Tagging:** POS tagging involves labeling words in a sentence with their corresponding part-of-speechcategories (noun, verb, adjective, etc.). This technique helps the NER model understand how words function in a sentence, improving entity recognition. SpaCy's POS tagging module is used in this project to analyze sentence structure and enhance NER accuracy. POS tagging considers features like previous and next words, word capitalization, and syntactic position to classify entities correctly.

Table -1: Named Entities with Entity tags

Named Entity	NE Tag
Person	per
Organization	org
Location	loc
Time	time
Geographical	geo
Geopolitical	gpe
Artifact	art
Natural phenomenon	nat

III. MACHINE LEARNING APPROACHES TO NAMES ENTITY RECOGNITION

For the classification and recognition of named entities, various machine learning techniques such as Hidden Markov Model (HMM), Conditional Random Field (CRF), Decision Tree, Support Vector Machine (SVM) are used.

- CRF based NER framework: Conditional random fields are a group of models that are best suited to predict contextual tasks. For labeling, Conditional Random Field is used. It is typically used for sequence labeling or parsing information, for example, processing of language and CRFs Named Entity Recognition for the POS labeling. For named object recognition activities, CRFs function well. For CRFs, characteristics can be used. For starters, on the lookout i.e. capitalization, attachments. CRFs are used to forecast sequences.

Denote x as the sequence of input states, i.e. the words of a sentence

$x = (x1,\ldots\ldots\ldots,xm)$

y as the output states, i.e. the named entity tags.

$y= (y1,\ldots\ldots\ldots,ym)$

For a conditional random field, we model a conditional probability

$\rho(y1,\ldots\ldots,ym| x1,\ldots\ldots xm)$

Define this by feature map

$\phi(x1,\ldots\ldots,xm,y1,\ldots\ldots ym) \epsilon Rd$

that maps an entire sequence of inputs x together with entire sequence y to some d-dimensional feature vector. Then model the probability with the parameter vector like a log-linear model. This penalizes the model complexity and is known as regularization.

$\omega \epsilon Rd$

4.1Training CRF

To train the CRF model, this project employs L-BFGS (Limited-memory Broyden–Fletcher–Goldfarb–Shanno) optimization algorithm, which efficiently handles large datasets with limited memory usage.

- Training Process: The training set consists of labeled sentences, where each word is tagged with its corresponding entity class.
- Regularization: To prevent overfitting, Elastic Net regularization (L1 + L2 penalty) is applied, ensuring model generalization on unseen data.

- Inference: Once trained, the model predicts named entity tags for test samples based on learned word associations and contextual dependencies.

While CRF models work well for structured and domain-specific text, they struggle with complex sentence structures. To address these limitations, deep learning-based NER models are used for enhanced accuracy.

4.2Bi-directional-LSTM-CRF Model

The bi-directional LSTM is a combination of two LSTMs, one running forward from ”right to left” and the other running backward from ”left to right.” Bidirectional long short term memory is used for the recognition of entities.

The two layers of LSTM are forward and backward layers. For capturing past dependencies forward layer is needed and the backward lstm layer is another layer storing future dependency. Entity Recognition is the most important technique for extracting, obtaining information, question answering, machine translation.

Character level vector concatenated as a word presentation with word embedding. Put it on the bidirectional LSTM first and the bidirectional LSTM is loaded into the CRF for label decoding.

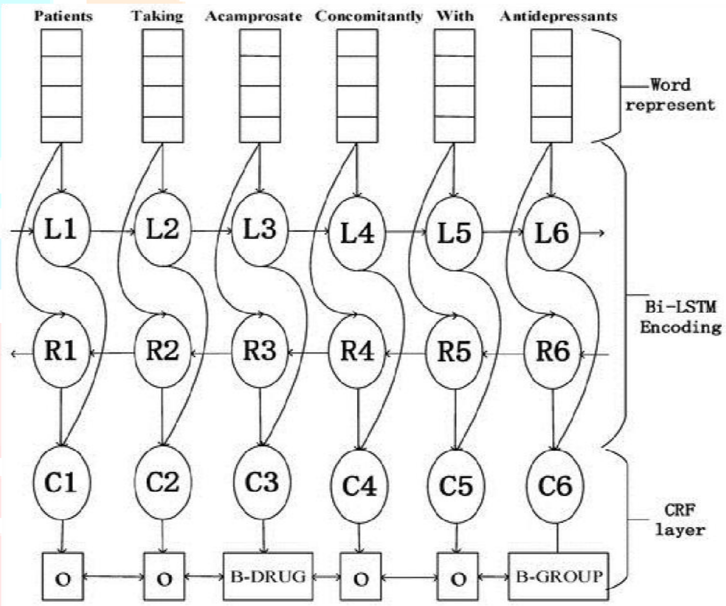

Fig. 2: Bi-LSTM-CRF Model

IV. PERFORMANCE METRICS

F1-score is used to measure performance metrics. It is the harmonic mean of precision and recall. **Precision** indicates how many of the predicted entities are actually correct, while **recall** measures how many of the actual entities were successfully identified by the model.

In this project, **precision, recall, and F1-score** are used as evaluation metrics to measure the model's efficiency. Since the dataset contains an imbalanced number of entity classes, accuracy alone is not a reliable metric. The table below presents the class-wise F1-score for named entity recognition (NER) in this project.

Table -2: Performance Metrics of NE tags

Performance Metrics

NE tag	Precision	Recall	f1- score
B-per	0.88	0.85	0.86
B-org	0.76	0.78	0.77
B-LOC	0.89	0.91	0.90
B-MISC	0.82	0.79	0.80
I-PER	0.86	0.88	0.87
I-ORG	0.74	0.80	0.77

V. CONCLUSION

The NER project successfully built a deep learning model for entity recognition in unstructured text. Using pre-trained models like BERT, it achieved high precision, recall, and F1-scores. Despite challenges like text ambiguity and domain adaptation, the project highlights the potential of deep learning in NER tasks. Future improvements include domain-specific models, multilingual support, and entity linking, benefiting industries like healthcare, legal, and customer support.

This project demonstrates the effectiveness of automated text understanding, paving the way for more advanced and scalable NER solutions in real-world applications.

VI. ACKNOWLEDGEMENT

I sincerely appreciate the resources and support that made this research possible. I am grateful for the tools, technologies, and knowledge that contributed to the successful completion of this work.

REFERENCES

[1]Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language
Understanding. arXiv preprint arXiv:1810.04805

[2]Shotaro Misawa, Motoki Taniguchi, Yasuhide Miura and Tomoko Ohkuma, "Character-based Bidirectional LSTM-CRF with
Words and characters for Japanese.
Named Entity Recognition,"Proceedings of the First Workshop on Subword and Character Level Models in NLP.

[3]Jason P.C. Chiu, Eric Nichols, "Named Entity Recognition with Bidirectional LSTM-CNNs,"IEEETransactions of the
Association for ComputationalLinguistics, vol. 4, pp. 357370, 2016

[4] Peng Sun, Xuezhen Yang, Xiaobing Zhao and ZhijuanWang. "An Overview of Named Entity Recognition,"International
Conference on Asian LanguageProcessing 2018.

[5]Qianjun Shuai, Runze Wang, Libiao Jin*,Long Pang"Research on Gender Recognition of Names Based onMachine Learning
Algorithms,"10th InternationalConference on Intelligent Human- Machine Systemsand Cybernetics 2018.

[6] Deepti Chopra, Nusrat Jahan, Sudha Morwal, "HindiNamed Entity Recognition Aggregating Rule basedHeuristic and Hidden

Markov Model ," International Journal of Information Sciences and Techniques (IJIST) Vol.2, No.6, November 2012.

[7] Devin Hoesen, Prosa Solusi, Cerdas Bandung, Ayu Purwarianti, "Investigating Bi-LSTM and CRF with POS Tag Embedding

for Indonesian Named Entity Tagger," International Conference on Asian Language Processing (IALP) 2018.

[8]Shotaro Misawa, Motoki Taniguchi, Yasuhide Miura and Tomoko Ohkuma, "Character-basedBidirectional LSTM-CRF with

words andcharacters for Japanese Named Entity Recognition,"Proceedings of the First Workshop on Subword and Character

Level Models in NLP.

[9]Xishuang Dong, Lijun Qian, Qiubin Yu, Jinfeng Yang, "A Multiclass Classification Method Based on Deep Learning for Named

Entity Recognition in Electronic Medical Records,"IEEE 2016.