Tejas Wani (65)
Aditya joshi (24)
Omkar Nandgaonkar (38)

Named Entity Recognition and POS Tagging for Indian Languages

Abstract: This project seeks to address the growing need for robust natural language processing (NLP) tools in the context of Indian languages. Specifically, it aims to create an integrated system capable of performing Named Entity Recognition (NER) and Part-of-Speech (POS) tagging for Indian languages. Indian languages present unique linguistic challenges, with rich morphology and a wide variety of entity types. The project's primary objective is to develop a versatile tool that enhances text understanding and information extraction in this diverse linguistic landscape.

Methodologies:

1. Data Collection:

- Diverse Dataset: The project will begin by collecting a comprehensive and diverse dataset of text written in Indian languages. This dataset will include languages such as Hindi, Tamil, Bengali, and more. The diversity of the dataset is essential for training the NER and POS models, as it will help account for variations in language structures, dialects, and writing styles.
- Annotation: To facilitate model training and evaluation, the dataset will be meticulously annotated. Named entities will be labeled, and words will be tagged with their respective part-of-speech categories. This annotated data forms the foundation for model development.

2. NER and POS Model Development:

- Customized Models: Machine learning models for NER and POS tagging will be developed using suitable NLP libraries and techniques. The critical aspect of model development is customization to accommodate the linguistic intricacies of Indian languages. This customization may involve adapting existing models to the specific characteristics of these languages.
- Morphological Analysis: Given the rich morphology of Indian languages, the NER and POS models will be fine-tuned to handle the complexities of word formation and inflection. This step is vital for accurate recognition and tagging.

- Language-Specific Features: The models will be designed to take advantage of language-specific features, such as character-level information and contextual cues that are particularly relevant to Indian languages.

3. Multilingual Support:

- Language Variations: Indian languages exhibit significant variations in terms of vocabulary, grammar, and entity types. To account for these variations, the NER and POS models will be extended to support multiple Indian languages.
- Language Identification: The system will include language identification capabilities to automatically detect the language of the input text, ensuring that the appropriate NER and POS models are applied.

4. Evaluation:

- Standard Metrics: The performance of the NER and POS models will be evaluated using standard NLP evaluation metrics such as precision, recall, F1-score, and accuracy. These metrics will gauge the models' ability to accurately recognize named entities and assign correct part-of-speech tags.
- Fine-Tuning: The project will involve iterative fine-tuning of the models based on evaluation results. This fine-tuning process aims to optimize the models' accuracy and effectiveness.

5. Integration:

- User Interface: The final output of this project will be a user-friendly interface or application. This interface will allow users to input text in Indian languages and receive real-time NER and POS tagging results. It is designed to be accessible to a broad audience, including non-technical users.

6. User Feedback and Improvement:

- Feedback Collection: To ensure the tool's accuracy and usability, feedback will be actively solicited from users, particularly from native speakers of Indian languages. User feedback is invaluable in identifying areas for improvement.
- Continuous Development: The project follows an iterative development approach. Feedback and insights collected from users will be used to update and expand the tool continually. This approach ensures that the tool remains relevant and continually improves its capabilities.

Technical Terms:

- Named Entity Recognition (NER): NER is a technique in NLP used to identify and classify named entities in text, such as names of people, places, organizations, dates, and more.
- Part-of-Speech (POS) Tagging: POS tagging involves assigning grammatical categories, such as nouns, verbs, adjectives, and adverbs, to individual words in a text.
- Morphology: Morphology deals with the structure and formation of words in a language, encompassing aspects like word roots, affixes, and inflections.
- Multilingual NLP: Multilingual NLP refers to NLP techniques that work effectively across multiple languages.
- Fine-Tuning: Fine-tuning is the process of adjusting and optimizing pre-trained models for specific tasks or domains.
- Linguistic Variation: Linguistic variation pertains to differences in language structure, vocabulary, and usage across different regions and dialects.

Use Case:

Imagine a scenario in which an Indian government agency seeks to engage with citizens across diverse linguistic regions. The NER and POS tagging tool developed in this project could be deployed to process and understand public feedback submitted in various Indian languages. By analyzing sentiment, extracting key information, and gaining insights into regional concerns, the agency could make data-driven decisions that cater to the specific linguistic and cultural contexts of its citizens. This tool is versatile and has the potential to benefit various domains where Indian language processing is essential, including government, media, and communication.

Conclusion:

In conclusion, this project addresses the pressing need for NER and POS tagging tools in the context of Indian languages. The development of models that can accommodate the rich linguistic diversity of these languages is a significant step forward in empowering users to better understand and extract information from texts in their native languages. The project also contributes to the ongoing development of NLP resources for underrepresented languages, which is an important aspect of promoting linguistic diversity and inclusion in the digital age. The iterative approach, guided by user feedback, ensures that the tool remains a valuable asset in the field of Indian language processing.