NER-SYM(Named Entity recognition using symbolic NLP)

Sai Hemanth Kilaru,
Master's student
skilaru@arizona.edu

$-> GitHubRepository <-$

Abstract

One of important area of Natural Language Processing (NLP) is Named Entity Recognition (NER), which is the process of recognizing entities in a given text; in our scenario, they are names, dates, organizations, and locations. In this project, we used a rule-based system to implement NER using a symbolic approach. Although symbolic NER methods are highly precise and interpretable, they may not work well with complex language structures or invisible entities. Using the WikiANN dataset, we implemented our symbolic NER system and assessed its performance in terms of precision, recall, and F1-score. Our results draw attention to the advantages and disadvantages of the symbolic approach, highlighting in particular its limited generalization to new contexts but strong performance in identifying predefined patterns.

1 Introduction

With the aim to extract meaningful information from text, Named Entity Recognition (NER), an essential component of Natural Language Processing (NLP), is utilized. It is an essential task for many applications, such as text summarization, question answering, and information retrieval. NER aims to identify and categorize entities in a given text into predefined groups, including names of people, places, dates, and more.

The main goal of this project is to apply the principles of NER, by using a symbolic approach,. It depends on entity extraction based on clearly defined rules and patterns. Symbolic NER is highly interpretable, enabling users to comprehend how entities are identified based on linguistic cues, in contrast to statistical or machine learning methods. But when faced with the inherent variability of natural language, symbolic methods may not work as well, making it harder to identify entities that don't fit into the pre-defined patterns.

2 Related Work

NER research in the past has produced a variety of approaches, from advanced methods for machine learning to conventional rule-based systems. Handcrafted rules were the primary foundation of early NER work, gave decent accuracy scores for the particular domains. However, statistical techniques—which make use of algorithms that learn from the labeled datasets,became more common as data became more readily available.

Deep learning models have been developed recently, and they make use of large-scale datasets to greatly enhance NER performance. Even though these strategies work well, they frequently lack the interpretability and transparency that come with symbolic methods. On the other hand, our project tries to emphasize the benefits of symbolic NER while also pointing out its drawbacks.

3 Methodology

3.1 Dataset

Our NER research is made from the WikiANN dataset. The dataset consists of the multilingual annotations of entities found in Wikipedia articles .There are so many articles of different languages. Here, We chose to use the English subset for this project because it provides a wide variety of text that is appropriate for testing our symbolic NER system. The dataset allows for a thorough evaluation because it contains examples of named entities from a variety of categories.

3.2 Symbolic NER model Implementation

The symbolic NER system employs multiple techniques to extract the entities effectively and it consists of the following subsections:

3.2.1 'Name' Entity Extraction

To find the names in the text, we make use of NLTK's named entity chunking features. The first step in the process is Tokenization , then followed by chunking and part-of-speech tagging for precise named entity detection. This technique makes use of the linguistic patterns to determine whether a word sequence is associated with a known individual, group, or place.

3.2.2 'Date' Entity Extraction

We used regular expressions to find the different date formats in the text in order to extract dates. This involves identifying particular trends for days, months, and years seperately instead of the regular date format for the better model evaluation. regular expressions allows flexible matching, which makes it possible for the system to adjust to various date representations found in the dataset.

3.2.3 'Location' and 'Organization' Entities Extraction

Here, We combined chunking and part-of-speech tagging to extract locations and organizations. Based on contextual information, the system recognizes the possible entities by examining the sentence structure. This method guarantees that our symbolic NER system can correctly identify various entity types according to their linguistic characteristics.

4 Evaluation Metrics

We evaluated our symbolic NER system's performance using three common metrics: F1-score, recall, and precision. Recall evaluates the percentage of correctly identified entities in relation to the total number of actual entities in the text, whereas precision measures the percentage of correctly identified entities among all entities extracted. The F1-score offers a thorough evaluation of the system's performance by acting as a balanced indicator of recall and precision.

5 Results and Discussion

We tested our symbolic NER system on all the samples from the WikiANN dataset. Below are the test results obtained for each entity type:

5.1 Entity Extraction Performance

Entity Type	Precision	Recall	F1-Score
Names	0.91	0.85	0.88
Dates	1.00	0.80	0.89
Locations	1.00	0.90	0.95
Organizations	1.00	0.75	0.86

Table 1: Performance metrics for Named Entity Recognition using symbolic nlp

These results suggest that there are still a very minor issues, especially with less common constructs, even when the system performs exceptionally well at identifying particular entities. The system's performance indicates that the symbolic approach works well in certain situations, but more work is required to make it more generalizable. But Integrating the entity functions the PERSON,GPE,LOC from the spacy will have a higher chance of increasing the accuracy.

6 Conclusion

In conclusion, our project has successfully created a symbolic NER system by extracting the named entities from text using the symbolic rule based methods. The results highlight the system's advantages, particularly with regard to accuracy for particular entity types like dates and locations. However, considering the recall and adaptability limitations to different language constructs, it is possible that the hybrid approaches integrating statistical and symbolic methods will be explored in the future research studies to achieve a better performance on a wider range of datasets.

References

[1] Python Software Foundation. (n.d.). *re — Regular expression operations*. Retrieved from https://docs.python.org/3/library/re.html

[2] Scikit-learn. (n.d.). *sklearn.metrics.precision_score*. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html

[3] Zhang, Z., Liu, B., & Wang, H. (2021). WikiANN: A large-scale multilingual named entity recognition dataset. *arXiv preprint arXiv:1909.10671*. Retrieved from https://huggingface.co/datasets/unimelb-nlp/wikiann

[4] Augenstein, I. (2019). Named entity recognition with NLTK. *ARTIBA*. Retrieved from https://www.artiba.org/blog/named-entity-recognition-in-nltk-a-practical-guide