Research Report

Dev Goyal

May 2023

1 Question 1

Named Entity Recognition (NER) is an important task in Natural Language Processing (NLP) that involves identifying and categorizing named entities in text documents. There are different methodologies for NER, each with its own advantages and drawbacks.

Rule-based NER entails creating a set of rules that define patterns and characteristics of named entities, which are then applied to the text for identification and categorization. The advantage of rule-based NER is its high accuracy and the ability to customize it for specific domains. However, developing and maintaining the rules can be time-consuming and requires expertise. Moreover, rule-based NER may struggle to identify named entities that don't fit the predefined rules.

Supervised learning for NER involves training a machine learning model on labeled data, where the model learns to identify and categorize named entities based on text features. Supervised learning offers high accuracy and the ability to handle named entities that don't conform to predefined rules. However, it demands a substantial amount of labeled data for training, which can be both time-consuming and expensive to acquire. Additionally, the model may not perform well on text documents that differ significantly from the training data.

Unsupervised learning for NER focuses on identifying patterns and relationships in text without relying on labeled data. This approach is useful for recognizing named entities that don't fit predefined rules or aren't present in the training data. However, unsupervised learning tends to be less accurate than supervised learning and may require additional manual processing to categorize the identified named entities.

To summarize, each NER methodology has its own strengths and weaknesses. Rule-based NER is highly accurate but demands domain expertise and may struggle with nonconforming named entities. Supervised learning offers high accuracy and flexibility but requires a large amount of labeled data. Unsupervised learning can handle diverse named entities but may be less accurate and require additional manual intervention. The choice of methodology depends on the specific NER task requirements and available resources.

2 Question 2

So in their paper "Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition" Minsoo Cho, Jihwan Ha, Chihyun Park, Sanghyun Park, discuss the bi-LSTM-CRF model for biomedical NER. They proposed passing every word of eevry sentrnce through a CNN and bi-LSTM model to get character level embeddings and then combine those with some pre trained word embeddings for biomedical words to get the 200 dimensional input for the main bi-LSTM-CRF model. After, reading through the paper I planned to replicate the results of their experiment on my own local machine by using tensorflow and python. The only constraint I faced was that I was bounded by the power of my own local machine and even after letting the code run overnight I got nowhere near close to converting every word of every sentence to their corresponding vector. This led me to reduce the size of the dataset to just 100 sentences with 2359 word different kinds of words and after going through