CS6301 Project Proposal
A Comparative Study of different approaches for Named Entity Recognition

Akshay Ramanujam Ranganathan
Chirag Rajesh Mukkatira
Meghana Spurthi Maadugundu
Siddhant Suresh Medar

1. Task

Introduction:
This project proposal aims to conduct a comparative study of various approaches for Named Entity Recognition (NER) in the context of Information Extraction. NER is a critical task in Natural Language Processing (NLP) that involves identifying and categorizing entities in text, such as people's names, locations, organizations, and time expressions. The main goal of NER is to tag a sequence of words with a label representing the type of entity the word belongs to.

Background:
Information Extraction involves extracting structured data from unstructured text, which is becoming increasingly important due to the abundance of text data available online. NER is an essential step in this process, as it helps to identify key elements in a text and sort unstructured data to detect important information.

Approaches:
Multiple approaches can be used for NER, such as rule-based, machine learning, and deep learning. Rule-based approaches use handcrafted rules to identify entities, whereas machine learning approaches learn from data to automatically identify entities. Deep learning approaches involve training neural networks to perform NER. Each approach has its strengths and weaknesses, and this project aims to compare their performance.

Objectives:
The main objectives of this project are:

- To identify the strengths and weaknesses of different NER approaches.
- To compare the performance of different NER approaches on various datasets.
- To investigate the impact of different factors on NER performance, such as dataset size and complexity.

- To propose guidelines for selecting the most suitable NER approach for a given task.

2. Data

The WikiNER dataset is a labeled dataset created by Nothman et al. in 2013 for Named Entity Recognition (NER) task, which involves identifying entities such as people, organizations, and locations in a text. The dataset consists of 7200 labeled Wikipedia articles across nine languages: English, German, French, Polish, Italian, Spanish, Dutch, Portuguese, and Russian. The English portion of the dataset contains 4853 labeled Wikipedia articles. The dataset is useful for NER tasks, as it contains a variety of entity types and has been labeled by human annotators.

The WikiNER dataset is a suitable choice for the proposed project, as it contains labeled data in the English language and has a large number of labeled documents. Additionally, the dataset contains entities from various domains and is suitable for training models that can recognize entities from multiple domains thus making it more generalized for real world applications. The dataset has been widely used in the research community for developing and evaluating NER models.

In the event that the dataset available is inadequate or does not yield satisfactory results, our approach is to combine the WikiNER dataset with other datasets like NERP, WNUT 2016, CoNLL 2003, and NERGrit. Our aim is to create a single, comprehensive dataset by merging several domain-specific NER datasets, which will enhance the model's ability to perform well in practical applications.

3. Methodology

The following methodology is followed for this project

Data collection: The first step in conducting a comparative study of various approaches for Named Entity Recognition is to collect suitable datasets. In this proposal, we will primarily use the WikiNER dataset. To address insufficient or underperforming datasets, we plan to merge WikiNER with other domain-specific NER datasets, like NERP, WNUT 2016, CoNLL 2003, and NERGrit. This approach will create a larger, more comprehensive dataset that can improve the model's performance and ability to generalize for real-world applications.

Preprocessing: The collected data will undergo preprocessing and augmentation to create a larger dataset that facilitates better generalization by the model.

Feature engineering: The next step is to extract features from the preprocessed data, which will be used to train and evaluate the NER models. We will use traditional features like part-of-speech (POS) tags, word embeddings, and character-level features, as well as more recent features like contextualized word representations and attention mechanisms.

Model training: We will train several NER models using various approaches, such as rule-based models, sequence labeling models like Conditional Random Fields (CRF), and neural network-based models like Recurrent Neural Networks (RNN) and Transformers. We will use the training data to optimize the models' parameters and fine-tune their hyperparameters.

Evaluation: We will evaluate the trained models using standard metrics like Precision, Recall, and F1-score. We will also perform a comparative analysis of the models' performance and identify the strengths and weaknesses of each approach.

Interpretation: After completing the experiments, we will analyze the results and draw conclusions on the effectiveness of the different NER approaches used. Based on our findings, we will provide recommendations and suggest ways to improve the state-of-the-art NER models for Information Extraction.

Example of an NER task: "Yesterday, the CEO of Microsoft, Satya Nadella, announced the opening of a new research center in Paris, France."

Tagged entities:
Organization: Microsoft
Person: Satya Nadella (CEO of Microsoft)
Location: Paris, France

4.Experiments and Analysis

Experimentation:
We will begin by training a baseline model using traditional machine learning techniques such as Conditional Random Fields (CRF) or Support Vector Machines (SVM) to set a benchmark for the NER task. Then, we will explore various deep learning models such as Bidirectional LSTM,

CNN-LSTM, and Transformer-based models for the NER task, as these models have shown state-of-the-art performance in various NLP tasks.

Analysis:

To compare the performance of different approaches, we will use standard evaluation metrics such as Precision, Recall, and F1-score, as well as calculate the accuracy and the time taken by each model to make predictions.

We will conduct a statistical analysis of the results obtained to determine if there is a significant difference in performance between the models. Finally, we will interpret the results and draw conclusions about the effectiveness of the various approaches for NER. We will also suggest recommendations and derived conclusions for building SOTA NER task model for Information Extraction.