CS535 - Natural Language Processing
Entities identification and their relation in Urdu Language

Muhammad Hasnain Khan
i19-1255

1 Problem Statement

NER and relation extraction of identified entities is a very tedious task in natural language and this task becomes more difficult when it comes to low resource language like Urdu because of the non-availability of the enough linguistic resources. In this project, I will work on the identification of named entities and their relation in Urdu Language.

2 Motivation

A lot of work has been done on **NER** for other languages but there can be found a few papers on NER for poor resource languages like Urdu. And moreover none of the literature can be found on relation extraction for Urdu language. There is a room of improvement as its a new direction for researchers and it will open door of research for many researchers. It will be beneficial for people to work in new environment with the help of available research content which is still lacking in this era of technology about this certain specific area of research.

3 Background

This project proposal is all about the identification of Named Entities from the text and their relation. Readers should have the basic knowledge of text processing and what is NER and how does it work.

4 Related Work

The research work on NER for Urdu language is in initial stage and the main reason is the non-availability of the standard NER resources to work on. To train a highly accurate model for NER, a large labeled NER dataset is needed [1]. There are a few dataset available online for Urdu NER but the size of the dataset is about a few thousand tagged entities. One of the dataset which is available publicly is IJCNLP-2008 tagged dataset. This dataset includes 40 thousands annotated words for twelve named entity classes. Second NER tagged dataset is annotated with four named entity classes [7]. It contains 31860 total words with 1526 named entities. These two datasets are available online and some useful information are extracted from the datasets which are given below.

Dataset	No. of Words	No. of Sentences	No. of NEs
Jahangir et al.,	31,860	1,315	1526
IJCNLP-2008	40408	1097	1115

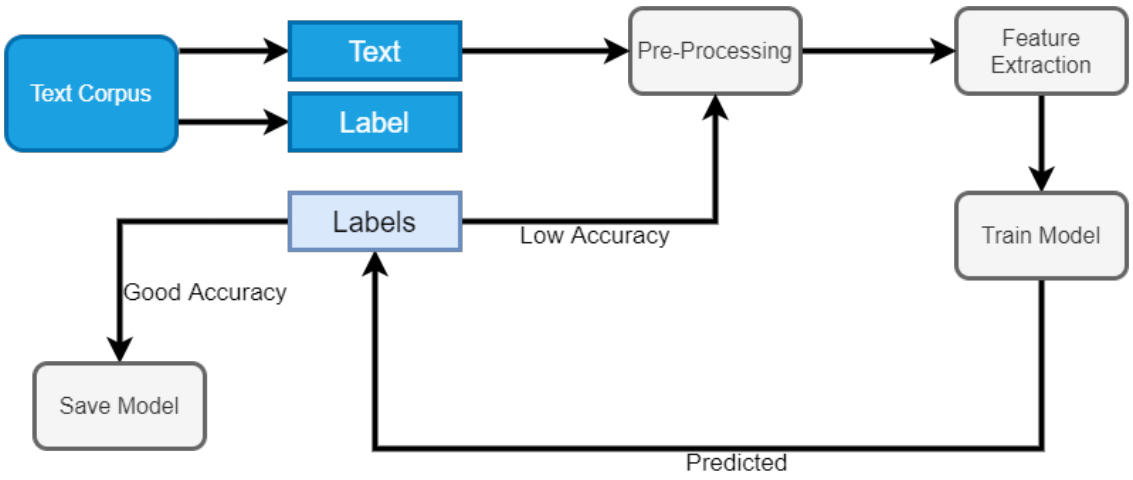
Figure 1: Details of Two NER dataset

Dataset	No. of Words	No. of Sentences	No. of NEs
Jahangir et al.,	31,860	1,315	1526
IJCNLP-2008	40408	1097	1115

Figure 2: Entity wise statistics

5 Proposed Work

A naive approach could be used to find these by looking at the noun phrases in text. We will be using supervised learning technique for NER and for that we will have to collect and label the data. Machine Learning model will be needed that will learn from the labeled data often referred as training data and model will be tuned on validation data which will be labeled. Different machine learning techniques will be used to identify NER but most likely Rule Based Learning or HMM with neural network will be used to identify NER. After entities identification, relationship between identified entities will be extracted. A lot of work has been done on this problem for different languages and same techniques will be used more or less to find out the relation between the identified entities. State diagram for the proposed system can be found below.



6 Evaluation Methodology

I will be evaluating my implementation with IIU research which is published on CLE and they achieved 81% accuracy. I will be comparing my results with their to evaluate the model. To validate the model, K-fold cross validation will also be used.

7 Hypothesis

- Extracted Features from the annotated dataset are relevant to this task.
- The trained model at the end will be the state-of-the-art model for NER.

8 Proposed Timeline

The tentative weekly timeline giving concrete milestones would be as follows.

- **October 12**: Proposal Document
- **October 17**:Gather NER Dataset
- **October 19**: Implementation of Deep learning model for NER
- **October 29**: Manually annotate relation extraction dataset
- **November 2**: Implementation of relation extraction system
- **November 7**: Model Evaluation
- **November 21**: Complete experiments.
- **November 27**: Writing report.
- **December 5**: Presentation and Final report due.

References

[1] W. Khan, A. Daud, J. A. Nasir, and T. Amjad, "A survey on the state-of-the-art machine learning models in the context of NLP," Kuwait journal of Science, vol. 43, pp.66-84, 2016.

[2] Singh, UmrinderPal, Vishal Goyal, and Gurpreet Singh Lehal. "Named entity recognition system for Urdu." In Proceedings of COLING 2012, pp. 2507-2518. 2012.

[3] Riaz, Kashif. "Rule-based named entity recognition in Urdu." In Proceedings of the 2010 named entities workshop, pp. 126-135. 2010.

[4] Tran, Thy Thy, Phong Le, and Sophia Ananiadou. "Revisiting Unsupervised Relation Extraction." arXiv preprint arXiv:2005.00087 (2020).

[5] Shahbazi, Hamed, Xiaoli Z. Fern, Reza Ghaeini, and Prasad Tadepalli. "Relation Extraction with Explanation." arXiv preprint arXiv:2005.14271 (2020).

[6] Gaut, Andrew, Tony Sun, Shirlyn Tang, Yuxin Huang, Jing Qian, Mai ElSherief, Jieyu Zhao et al. "Towards Understanding Gender Bias in Relation Extraction." arXiv preprint arXiv:1911.03642 (2019).

[7] Khan, W., A. Daud, J. A. Nasir, and T. Amjad. "Named entity dataset for urdu named entity recognition task." Organization 48 (2016): 282.