IT412: Natural Language Processing

Assignment 4 : Named Entity Recognition

Instructor: Prasenjit Majumder

Learning Outcome: At the end of this assignment you will learn about Named Entity Recognition and applying classification algorithms for Named Entity Recognition and evaluating them

1 Problem description

Named Entity Recognition (NER) is a process of recognizing information units like names, including person, organization and location names, and numeric expressions including time, date, money and percent expressions from unstructured text. NER is used in many fields of Natural Language Processing, and it can help to answer many real-world questions. The goal of this assignment is to develop practical and domain-independent techniques in order to detect named entities with high accuracy automatically using Conditional random field (CRF).

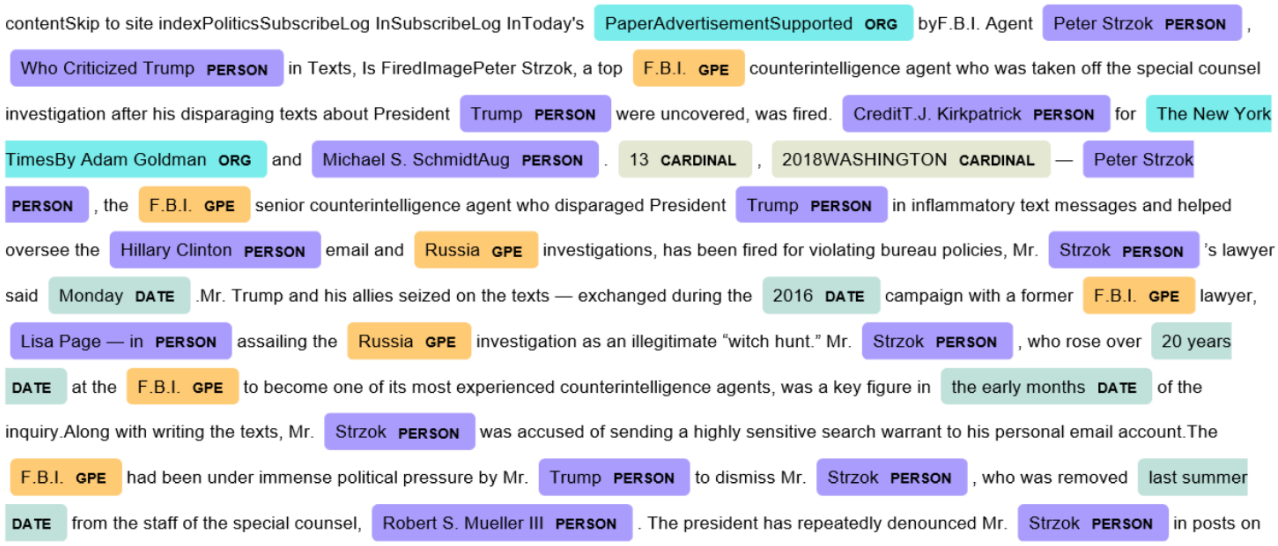

Figure 1: NER Example

1.1 Conditional Random Fields

Conditional Random Fields is a class of discriminative models best suited to prediction tasks where contextual information or state of the neighbors affect the current prediction. CRFs applications in named entity recognition, part of speech tagging, gene prediction, noise reduction and object detection problems, to name a few.

CRFs are used for predicting the sequences that use the contextual information to add information which will be used by the model to make a correct prediction. Below is the formula for CRF where y is the output variable and X is input sequence.

$$p(y|X,\lambda) = \frac{1}{Z(X)} exp \sum_{i=1}^{n} \sum_{j} \lambda_j f_j(X, i, y_{i-1}, y_i)$$

The output sequence is modeled as the normalized product of the feature function.

2 Implementation

2.1 Dataset

- GMB(Groningen Meaning Bank) corpus Which contains 47959 sentences containing 35178 different words.
- Download the dataset using the link: https://drive.google.com/file/d/19CtxJxsYs8hFMXpHkcmhD612OG86943e/view?usp=sharing

2.2 Exercise

1. Divide the dataset into 80:20 ratio
2. Represent the word using the following features: POS tag and word represented using one hot encoding
3. Apply Condition Random Field (CRF) and perform NER
4. Apply Stochastic Gradient Descent (SGD) and perform NER
5. Evaluate using Precision, Recall and F1 score. Report scores for each label as well as macro level.

3 References

- Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data
- NER using Multinomial Naive Bayes

3.1 Codes

- Implementation of NER using CRF

4 Submission

- Your notebook should contain the scores mentioned and proper comments for your approach
- The submission deadline for this assignment is **13th September 2021 at 11 PM**