CS 445 - Natural Language Processing

Project 4

Instructor: Reyyan Yeniterzi

Engincan Varan 25050

24.01.2021

Introduction 2

Results 3

Analysis 4

Future Works 5

Introduction

In this project, the aim is to create an effective NER tool for Turkish language using machine learning approaches. Here utilized the crfsuite model of sklearn. We had 2 files one of which includes Turkish paragraphs with some entity names, and the other has morphological analysis of the words in the paragraphs. Our aim is to create a feature map of the words and train a model that can give good results.

For feature mapping, we looked for some certain features such as (mapFeatures function in the codes) :

- Root (Stem)
- Part-of-Speech (POS)
- Proper Noun (PROP)
- Noun Case (NCS)
- Orthographic Case (OCS)
- All Inflectional Features (INF)
- Start of the Sentence (SS)

Also, we used a database of organizations to see improve our scores by adding another feature which checks whether the word is in the database of organizations or not.

- Is in Organization Database (INORG)

After defining the features and extract the labels from the given data we end up with 7 different labels for our words, which are:

- O (out of context)
- B-LOCATION (beginning of a location entity)
- I-LOCATION (inside of a location entity)
- B-ORGANIZATION (beginning of a organization entity)
- I-ORGANIZATION (inside of a organization entity)
- B-PERSON (beginning of a person entity)
- I-PERSON (inside of a person entity)

An example data type with all the features and the label:

```
{  'Stem': 'müzik',      'POS': 'Noun',        'PROP': False,
   'NCS': 'Nom',         'OCS': 'LC',          'INF': 'A3sg+Pnon+Nom',
   'BOS': True,          'INORG': False} --> O
```

The data is folded into 5 folds, each of them containing 2000 sentences. Therefore, while training the model and testing it, we used 5-fold cross validation and averaged the results of accuracy, f1, precision and recall.

Results

Finally, we created the model with different feature combinations (adding the features one by one to see the effects) Here are the results from the model:

Results for: Stem BOS Flat_Accuracy: 0.9498404755187663 Flat_F1_Score: 0.6162452968455749 Flat_Precision: 0.7406090322829657 Flat_Recall: 0.5289886961436592	**Results for: Stem POS BOS** Flat_Accuracy: 0.9509958062261678 Flat_F1_Score: 0.6304383299263348 Flat_Precision: 0.7403506217363869 Flat_Recall: 0.5502953544161597	**Results for: Stem POS BOS** Flat_Accuracy: 0.9509958062261678 Flat_F1_Score: 0.6304383299263348 Flat_Precision: 0.7403506217363869 Flat_Recall: 0.5502953544161597
Results for: Stem POS PROP BOS Flat_Accuracy: 0.9520129512537944 Flat_F1_Score: 0.6415555321490359 Flat_Precision: 0.7368059723799022 Flat_Recall: 0.5698045301776119	**Results for: Stem POS PROP NCS BOS** Flat_Accuracy: 0.9720137212384043 Flat_F1_Score: 0.7788600017220744 Flat_Precision: 0.8591605185049479 Flat_Recall: 0.7152766004811129	**Results for: Stem POS PROP NCS OCS BOS** Flat_Accuracy: 0.9721666719171731 Flat_F1_Score: 0.7816482114219917 Flat_Precision: 0.8599564901518695 Flat_Recall: 0.7185161466851903

Results for: Stem POS PROP NCS OCS INF BOS Flat_Accuracy: 0.9737891772606252 Flat_F1_Score: 0.7910719529407627 Flat_Precision: 0.8518346262471604 Flat_Recall: 0.7420261802240048	**Results for: Stem POS PROP NCS OCS INF INORG BOS** Flat_Accuracy: 0.9735269537012684 Flat_F1_Score: 0.7888265339999256 Flat_Precision: 0.850727165670272 Flat_Recall: 0.7387808388163293

See the html output for more details.

Note that, the dataset is not balanced, the number of labels are:

```
{   'B-LOCATION': 2506,
    'B-ORGANIZATION': 2402,
    'B-PERSON': 4568,
    'I-LOCATION': 343,
    'I-ORGANIZATION': 1587,
    'I-PERSON': 1927,
    'O': 152515     }
```

Therefore while evaluating the model, the accuracy score is not the best metric to evaluate.

Also, we omit the “O” label from the confusion matrix while evaluating the model.

Analysis

From the results we can clearly see that as we add more features to the model, in terms of morphological analysis tokens, we get better and better results. For example, only having “stem” and “BOS” (beginning of the sentence) features resulted in around 0.616 f1 score whereas having all the features (except “INORG”) resulted in 0.742 which is a significant increase. This is expected since the dataset is large enough to say that more features means a more accurate model.

After evaluating the model, we printed out the model transition feature to see what the model is doing in order to find the entities. In this case, these are the top and bottom transitions:

```
Top likely transitions:
I-ORGANIZATION -> I-ORGANIZATION 4.347910
B-ORGANIZATION -> I-ORGANIZATION 4.002563
I-LOCATION -> I-LOCATION 3.874740
B-PERSON -> I-PERSON 3.843507
B-LOCATION -> I-LOCATION 2.428487
I-PERSON -> I-PERSON 2.182171
I-ORGANIZATION -> O       1.858638
O      -> O       1.353891
B-ORGANIZATION -> O       1.298995
B-ORGANIZATION -> B-LOCATION 0.871935
I-LOCATION -> O       0.771517
B-PERSON -> O       0.768013
O      -> B-PERSON 0.733506
I-PERSON -> O       0.606989
O      -> B-LOCATION 0.377046
I-LOCATION -> B-LOCATION 0.187731
B-LOCATION -> O       0.155373
B-ORGANIZATION -> I-LOCATION 0.086881
O      -> B-ORGANIZATION -0.054770
I-ORGANIZATION -> B-ORGANIZATION -0.089100
```

```
Top unlikely transitions:
I-PERSON -> B-ORGANIZATION -0.970206
B-ORGANIZATION -> B-ORGANIZATION -0.989085
B-PERSON -> B-LOCATION -1.034096
B-ORGANIZATION -> I-PERSON -1.192168
I-LOCATION -> B-PERSON -1.387936
I-LOCATION -> B-ORGANIZATION -1.456096
I-PERSON -> B-PERSON -1.821215
I-ORGANIZATION -> I-PERSON -2.308405
B-LOCATION -> B-ORGANIZATION -2.374538
I-PERSON -> I-LOCATION -2.450532
I-LOCATION -> I-PERSON -2.485187
B-ORGANIZATION -> B-PERSON -2.596828
B-PERSON -> I-LOCATION -2.626294
B-LOCATION -> I-ORGANIZATION -2.724954
O      -> I-PERSON -3.026983
B-LOCATION -> B-PERSON -3.043153
O      -> I-ORGANIZATION -3.483420
B-PERSON -> I-ORGANIZATION -3.602243
O      -> I-LOCATION -3.614609
B-LOCATION -> I-PERSON -3.713954
```

From these tables, we see that the model thinks after the beginning of an organization name it is likely that it will be followed by another name which belongs to the same organization, yet it is unlikely that a beginning of an organization name will be followed by another beginning of an organization name. If we think about it, it seems very accurate to say that, organization names commonly have more than one word named entities, so the beginning of the organization name will be followed by the rest of the name. These tables can be analyzed like this to have an idea about what the model is thinking and applying.

Future Works

- We can always improve the scores by adding more features to our feature mapping as basic as "EOS" which is the end of the sentence feature for each word.
- We can provide more gazetteers for our organization labels.
- We can provide more gazetteers for every label we have.
- We can fine-tune the model using RandomizedSearchCV, which is implemented in the codes but not tested since it was not in the scope of the project.