Sabancı Üniversitesi

CS 445 Natural Language Processing

Project 4: Named Entity Recognition

Due Date: January 24, 23:55

This is the final version of the project. Please make sure to do everything described here.

In this project, you will develop an effective NER tool for Turkish using Machine Learning approaches. Use the CRF implementation available within Python sklearn, called crfsuite. You will try different features with CRF and calculate the performance scores. You will write your findings, results and interpretations into a report and submit that as well.

Dataset:

You are provided with a Turkish dataset which has been labeled for PERSON, LOCATION and ORGANIZATION. The dataset is already available at SUCourse, please download it from there. Further details of this dataset is provided in the papers [1, 2, 3].

You are responsible for splitting the dataset into 5 folds. You will iterate over sentences (lines) one by one and put the first one to first fold, the second one to second fold and so on. There are 10000 lines. So each fold should have 2000 sentences. At the end you are going to report your scores which are averaged over these 5 folds.

In addition to the original data, you are also provided with the morphological analysis of the words. You can use this data for constructing additional features.

Some Relevant Literature:

The following literature contain a list of useful features to identify Named Entities: [2, 3]. You are provided with a smaller dataset used in these papers. Since the datasets are different, your results are not directly comparable. However, as you add more features your performance should improve similar to the papers. If you have any questions regarding these papers, please email to the instructor (reyyan@sabanciuniv.edu).

You are expected to use the following features.

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

- Start of the Sentence (SS)

For more information on these features please check [2, 3].

In addition to the above features, please use a gazetteer based feature. You can use one of the gazetteers from your Project 01. You will implement this feature as an indicator feature. If there is a match, it will return 1, otherwise it will return 0.

Implementation Part:

In this project, you are expected to use Google colab. You will do your implementations on ***project04.ipynb*** and submit it with the expected outputs.

In [2, 3], the authors applied entity level evaluation. In order to keep things simpler for you we will accept token/tag based evaluation. So you will report your precision, recall and F1 for B-PERSON, I-PERSON etc. in your report.

Make sure that your *project04.ipynb* is well commented. After running all the cells, export the html output. You are going to submit both the ipynb and html with outputs.

You can use the popular/standard python packages. In case you are not sure of a particular library, please ask the instructor or TA.

You are expected to implement this project at your own. Your scripts will be analyzed by using state-of-the-art tools for any type of plagiarism.

Report:

In your report, you are going to summarize your approach and your findings. Discuss what is working and what is not. You do not need to write an introduction or related work section like in the papers. But you need to describe your features and their results (Precision, Recall and F-Measure) in your report. You need to interpret your results.

You are expected to write this report at your own. Your report will be analyzed with Turnitin.

All ipynb and html files should be under the same directory (named as your student ID, only the 5 digit numbers), which you will zip and submit. You will also submit your report. Your report is not going to be in the zip file. You will submit two files: (1) zip file and (2) report in pdf.

Submission Instructions:

- You will submit this homework via SUCourse.
- Please check the slides for the late submission policy.
- You can resubmit your homework (until the deadline) if you need to.
- Please read this document again before submitting.

- After submitting, you should download your submission to a different path to double check whether everything is in order.
- Please do your assignment individually, do not copy from a friend or the Internet. Plagiarized assignments will receive -100.

References

[1] Gökhan Tür, Dilek Z. Hakkani-Tür, and Kemal Oflazer. (2003). A statistical information extraction system for Turkish. In Natural Language Engineering, pages 181–210.

[2] Reyyan Yeniterzi. (2011). Exploiting morphology in Turkish named entity recognition system. In *Proceedings of the ACL 2011 Student Session*, Portland, OR, USA.

[3] Gökhan Akın, Gülşen Eryiğit. (2012) Initial explorations on using CRFs for Turkish Named Entity Recognition, COLING.