Boğaziçi University

Institute for Data Science and Artificial Intelligence

DSAI 545 – Natural Language Processing

Spring 2023-2024

Term Project – Cross-Domain Named Entity Recognition

Instructor: Dr. Şaziye Betül Özateş

21.04.2024

Project Submission Deadline: 03.06.2024, 23:59 GMT+3

All project related documents will be sent to saziye.ozates@bogazici.edu.tr

Project Description

Cross-domain Named Entity Recognition (NER) refers to the task of extracting named entities (such as persons, organizations, locations, etc.) from text across different domains or topics.

In NER, the goal is to identify and classify these named entities into predefined categories. Cross-domain NER extends this task by addressing the challenge of recognizing named entities in text from diverse domains or fields, where the language, vocabulary, and types of named entities may vary significantly.

Developing a cross-domain NER system thus requires building models that can transfer knowledge from one domain to another, enabling effective named entity recognition across a wide range of texts, even those with different styles, vocabularies, and linguistic nuances.

In this project, you are asked to develop a NER solution for judicial 'Ruznamçe' registers written in the 17th-18th centuries.

Ruznamçe registers were daily records kept by governors, detailing events, decisions, and other administrative matters. These registers were crucial for maintaining records of governance, taxation, legal proceedings, and other important matters.

Since the target domain is a specific one (judicial documents about legal matters), a NER model trained on a general entity recognition dataset would perform poorly without transfer learning.

Now, the fun part ☺

What you will be given:

Supervised training data: A NER dataset for Ottoman Turkish which includes annotated sentences from different issues of Servet-i Funun journal (published between 1896-1901). The dataset has 462 sentences from various topics including literature, science, daily life, and world news. There are two types of named entities tagged in the sentences: PERSON and LOCATION. You are going to use this dataset to train your NER system.

Additional data: Raw texts from judicial Ruznamçe registers. Usage of these texts are optional. For example, if you decide to utilize an LLM-based model, you may want to use the raw data to 'further pretrain' the employed LLM. Or, you can adapt the vocabulary of your system to the target domain by utilizing these texts. Search ways of utilizing unsupervised learning approaches.

Test data: A small, annotated subset of the judicial Ruznamçe registers. Named entities annotated in the test set are the same with the ones in the training data: PERSON and LOCATION.

What is being asked of you:

Build a successful NER system for the target domain.

Main Challenges:

- The lexicon of the source domain (general texts from Servet-i Funun journal from late 19th century and the beginning of 20th century) and the target domain (judicial Ruznamçe registers from 17th and 18th centuries) differ greatly.

- Both the source and target domains are extremely low-resource. E.g., only 462 labeled sentences from the source domain, no labeled data from the target domain.

- Existing NLP models and tools for Turkish only covers modern Turkish, no NLP model for historical Turkish.

DO's:

- Cite every code, every resource, everything you use.
- There is no restriction in the methodology of the project. You can use any NLP technic and public resource in order to enhance the performance of your system.
- You can use Google Colab to develop your models.

DON'T's:

- Do not cheat ☺

What to submit:

- A heavily commented source code of your system.
- A README file describing the requirements and steps to run your system.
- A project report that explains methodology of your NER system in DETAIL.
    - Mandatory sections in the report:
        - Introduction
        - Model description
        - Additional resources (if any): Write name, citation, link, and a short description for any resource/code/system you utilized in your NER system.
        - Evaluation results: Report the precision, recall, and F1-score of your system on the test data.

Note: All of the above points will be taken into consideration in grading your project. Projects submitted after the submission deadline will not be graded.