NLP Assignment 1 - Named Entity Recognition

Description

Named entities are phrases that contain the names of persons, organizations, geo-locations, companies, facilities etc.

e.g., *U.N. official Ekeus (person) heads for Baghdad (geo-location)*

The task of WNUT-2016 concerns named entity recognition. We will concentrate on 10 fine-grained NER categories: **person, geo-location, company, facility, product, music artist, movie, sports team, tv show and other.** Dataset was extracted from tweets in English language.

The named entity tags have format **B-TYPE**, **I-TYPE**, **O**. The first word of the entity will have tag **B-TYPE**. **I-TYPE** means that the word is inside an entity of type "TYPE". Words with tag **O** do not belong to any entities.

For this assignment, you are offered training data, validation data and testing data (no entity tags). You need to use the training data for training deep learning model, the validation data for tuning hyper-parameters to improve performance of the model, the testing data for final evaluation. When the training process is done, you use the well-trained model to fill out the testing data with predicted named entity tags as submission. We will evaluate your submissions with ground truth.

Requirements

- **Python programming language only.**
- **You can use any machine learning library (TensorFlow, Keras, Pytorch, etc.).**
- You can refer to the snippets of code from open-source projects, tutorials, blogs, etc. Do not clone the entire projects directly. **Try to implement a model by yourself.**
- **Word embedding like Word2vec or Glove should be used in the project.**

Grading

We follow the definition of metrics introduced at CoNLL-2003 to measure the performance of the systems in terms of precision, recall and F1-score, where:

"precision is the percentage of named entities found by the learning system that are correct. Recall is the percentage of named entities present in the corpus that are found by the system. A named entity is correct only if it is an exact match of the corresponding entity in the data file."

Models are evaluated based on exact-match F1-score on the testing data.

• Completed Source Code (50%)

• F1-Score (20%)

• Report (30%)

Submission Rule

Please pack up your **source code**, **test-submit.txt** and **report** into a .zip file named as *<studentid>_hw1.zip*, and upload it to the ee-class system. The following files must be included in your submission:

- One Python source code *<studentid>.py*
- test-submit.txt: test-submit.txt with predicted entity tags. Please follow the same format as dev.txt and do not rearrange the lines.
- requirements.txt: The list of installed libraries in your Python environment. To create a requirements file, use command: *pip freeze > requirements.txt*
- *<studentid>_report.pdf*: assignment report. You can write your data preprocessing, model architecture, training process, evaluation scores or anything else you want.

Deadline: 2022/11/08 23:59

Key words: sequence labeling, NER