Homework for the "Deep Learning for NLP" module*

Artem Chernodub, Ph.D.
Ukrainian Catholic University, Faculty of Applied Sciences
Grammarly
chernodub@ucu.edu.ua

July 21, 2019

The goal of the homework is to develop a tool for Named Entity Recognition. You need to implement model *"Glove word embeddings + BiLSTM + Softmax"* for sequence labeling. Please, use the standard PyTorch example for the sequence labeling task "Sequence models and long-short term memory networks" as a basic code to start. Glove word embeddings can be downloaded here.

Please, add the following modifications:

1. Implement functionality to read and process NER 2003 English Shared Task data in CoNNL file format, data will be provided (10% of score).
2. Implement 3 strategies for loading the embeddings:
   (a) load the embeddings for original capitalization of words. If embedding for this word doesn't exists, associate it with *UNKNOWN* embedding (5% of score).
   (b) load the embeddings for lowercased capitalization of words. If embedding for this lowercased word doesn't exists, associate it with *UNKNOWN* embedding (5% of score).
   (c) load the embeddings for original capitalization of words. If embedding for this word doesn't exists, try to find the embedding for lowercased version and associate it to the word with original capitalization. Otherwise, associate it with *UNKNOWN* embedding (20% of score).

*Yeah, it's time to start working on it right now.

3. Implement training on batches (20% of score).

4. Implement the calculation of token-level *Precision / Recall / F1 / F0.5* scores for all classes in average. IMPORTANT! Please, implement "micro-average" approach. Don't use standard functions from scikit-learn or similar external packages (30% of score).

5. Provide the report the performances (F1 and F0.5 scores) on the dev / test subsets w.r.t epoch number during the training for the first 5 epochs for each strategy of loading the embeddings (10% of score).

The deadline is 5/08/2019, 9:00 AM CET 9:00 AM CET. The penalty for missing the deadline: up to one week it is 50%, more than one week it is 100% of scores.