Homeworks: Natural Language Processing

Simone Antonelli

Master degree in Computer Science
Sapienza, University of Rome

A. Y. 2019 - 2020

Homework 1: Named Entity Recognition

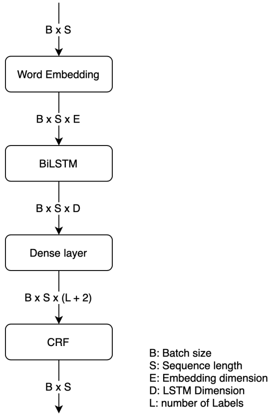
Named Entity Recognition (NER) task is a sequence labeling task that consists in identifying the categories to which words of a sentence belong:

- Organization (ORG);
- Location (LOC);
- Person (PER);
- Other (O).

Homework 1: Architecture

The adopted architecture to solve NER task is composed by:

- Word embedding layer;
- Bidirectional LSTM;
- Conditional Random Field.



Homework 1: Word Embedding

A word embedding is a learned representation for sentences where words that have similar meaning should have a similar representation.

The best performances of the model are obtained using fastText pretrained word embedding[1].

[1] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages.

Homework 1: BiLSTM

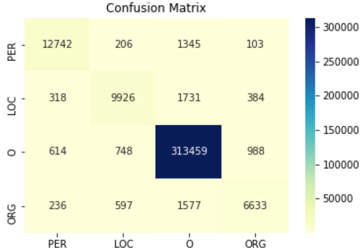
- Long short-term memories (LSTM) are a variant of the Recurrent Neural Networks (RNN)
- The Bidirectional LSTM is used to create the context of each token in the sentence leveraging two different LSTM to capture the left-context and the right-context
- Useful for the conditional random field layer

Homework 1: CRF

- Conditional probabilistic model used when there are dependencies across output labels
- In NER task to predict the label of each token is needed to make use of neighbor tag information
- It uses a score function to predict the best sequence of labels

Homework 1: Results

Due to the high imbalance of the dataset the metric used to evaluate the model is the macro f1-score.



Model	F1-score
BiLSTM + W + CRF	88%

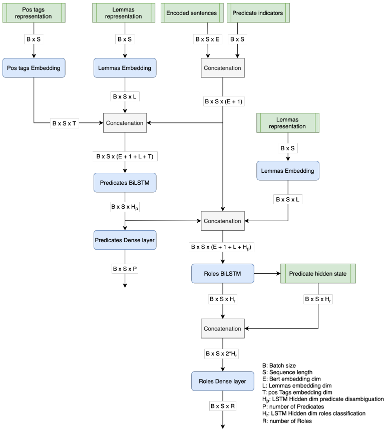
Homework 2: Semantic Role Labeling

Semantic Role Labeling is the task that extracts the predicate-argument structure of a sentence. It is divided in 4 steps:

1. Predicate identification
2. Predicate disambiguation
3. Argument identification
4. Argument classification

Homework 2: Architecture

Since the faced tasks seems like sequence labeling tasks, the main component of the model is the bidirectional LSTM.



Homework 2: Data preprocessing

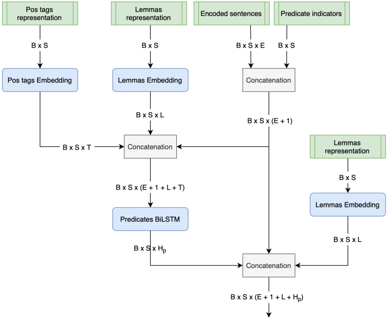
To represent the sentences is used BERT, a transformer-based model which gives as output the contextualized embeddings of each word of a given sentence.

The representation given by BERT is used to fed the model[2].
The pretrained model used in this homework is the simplest ('bert-base-cased') which has 12 layers and a hidden size equals to 768.

[2] Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling.

Homework 2: Encoding input

Further information about the sentences are provided by the given dataset, so those are leveraged to give to the model much information as possible on the sentences' tokens[3].



[3] Diego Marcheggiani, Anton Frolov, and Ivan Titov. 2017. A simple and accurate syntax-agnostic neural model for dependency-based semantic role

Homework 2: How the model works

- The model is fed with a sentence-predicate pair, so the sentences with many predicates are splitted online;
- The representation is given as input to the BiLSTM which deals with predicate disambiguation;
- The output is concatenated to the embeddings and then is given as input to the BiLSTM that faces role classification;
- Before doing role classification, to the hidden state of each token is concatenated with the hidden representation of the predicate in the sentence.

Homework 2: Results

Architecture model	Predicate disambiguation	Argument identification	Argument classification
BiLSTM $(W + P_i)$	-	84.48 %	76.45 %
BiLSTM $(B + P_i)$	-	91.30 %	82.50 %
BiLSTM $(B{+}P_i{+}T{+}L{+}P)$	-	**93.91 %**	**90.09 %**
Stacked BiLSTM $(B + P_i + O + L)$	**95.17 %**	**93.50 %**	**87.83 %**