Interuniversity Master's Degree in Artificial Intelligence

Natural Language Understanding – Academic Year 2022-2023

Practical Assignment 1 - Part-of-speech (PoS) Tagging

In this assignment, you will implement a neural PoS sequence labeling model to label the words of an input given sentence according their morphological information. Table 1 shows a simple example:

Inputs	Google	is	a	nice	search	engine	.
Outputs	PROPN	AUX	DET	ADJ	NOUN	NOUN	PUNCT

Table 1. Example of an input sentence and the ground-truth output, extracted from the test set of the UD_English_EWT dataset.

More particularly, you are asked to implement a neural model based on long short-term memory networks (LSTMs). We propose a default architecture. The basic model must receive as input word-level embeddings. Then, this representation will be fed to a word-level LSTM layer, followed by a dense layer, which will output the label for each token using a softmax activation function.

NOTE: Additionally, consider a variant of the model that in addition to word-level embeddings also uses, as input to the LSTM, character-level embeddings computed through a character-level LSTM (we will refer to these representations as character-level word representations). For each word, these two inputs (word-level embeddings and character-level word representations) should be then concatenated before being fed to the word-level LSTM. Without implementing the character-level functionality, the maximum score of this assignment will be 8 out of 10 points.

Figure 1 shows the high-level architecture of the network. You have liberty to modify the dimensions of the network, explore hyperparameters to improve the performance, and expand the base model.

The functionalities that must be supported are:

1. **Train a model,** and verify that converges successfully.
2. **Evaluate a trained model,** and report the tagging accuracy on both the validation/development and test sets that we indicate below.
3. **Once your model is trained, implement a function to compute the part-of-speech tags for given new, previously unseen sentences inputted by the user.**

Specifically, for this assignment you must train and evaluate your model on the English_EWT dataset from the https://universaldependencies.org/ (UD) collection, using the corresponding training, development and test files, represented in CoNLL-U format (the details about this format can be found at: https://universaldependencies.org/format.html). Here, we will use and learn to predict the morphological information stored in the UPOS column. In addition, we also ask you to train and evaluate the model in two other language datasets from UD, of your own choice. For simplicity, we will not consider input sentences longer than 128 words.

For more details about the assignment, you can also check the guide that accompanies this assignment.

Submission

- The assignment must include a short user manual that indicates how to run the code for training, evaluation, and generation of labels. It must also include a brief discussion of the implementation decisions, differences across the evaluated models that you might have explored, as well as an analysis of the performance across the evaluated datasets. The document must not exceed 3 pages, and it will be written using Calibri font style and a size of minimum 11pt.

- The assignment will be done in groups of 2 people.

- The assignment will be submitted to the virtual campus, on November 13 at 23:59 the latest. Only one member of the group will submit the assignment. Assignments submitted after the deadline will be automatically considered failed with a score of 0 out of 10 points.

- The assignment will be defended during the next weeks after the submission, during the laboratory hours. The slots will be made available by the professor at the time. Both members of the group must be present during the defense, otherwise the missing member(s) will fail the assignment with a score of 0 out of 10 points.

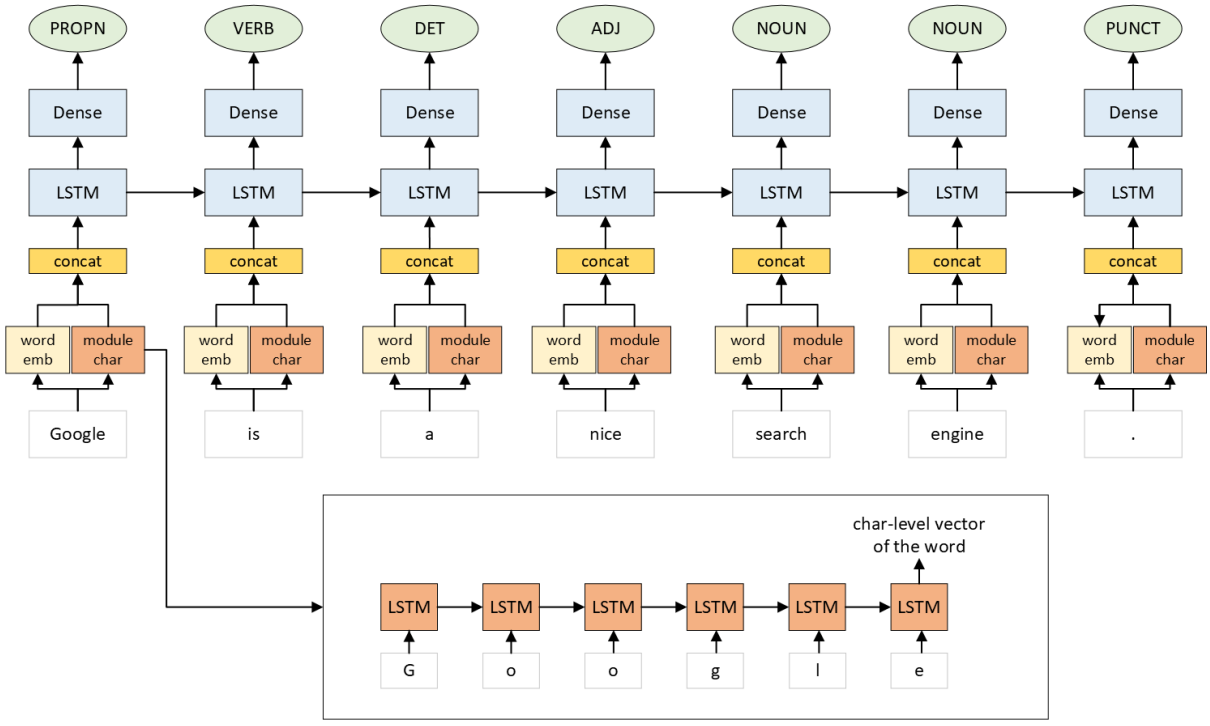

Figure 1: High-level architecture of the PoS tagging model