UNIVERSIDADE DA CORUÑA
USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
UniversidadeVigo

Interuniversity Master’s Degree in Artificial Intelligence

Natural Language Understanding – Academic Year 2025-2026

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 a given input sentence according to 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.

Specifically, your task involves the implementation of a neural model utilizing long short-term memory networks (LSTMs). We provide a suggested default architecture. The fundamental model should take word-level embeddings as its input. Subsequently, this representation will be passed through a word-level LSTM layer, followed by a dense layer that assigns a label to each token using a softmax activation function.

Figure 1 illustrates the high-level structure of the network. You are granted the flexibility to adjust the network's dimensions, experiment with hyperparameters for enhanced performance, and extend the foundational model.

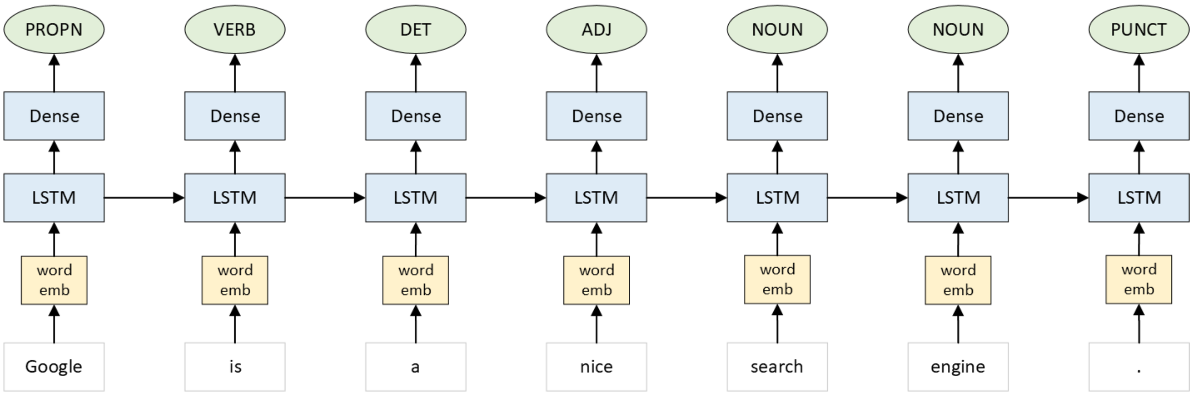

Figure 1: High-level architecture of the PoS tagging model

The functionalities that must be supported are:

1. **Train a model and ensure it converges successfully.**
2. **Evaluate the trained model** and provide tagging accuracy metrics for both the validation/development and test sets specified below.
3. **Develop a function to compute part-of-speech tags for newly input sentences** by the user that have not been seen before.

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. Furthermore, we request that you train and evaluate the model on two other language datasets from UD of your choosing. To simplify the process, sentences longer than 128 words can be preprocessed and removed during the training and evaluation of your models.

For further guidance on this assignment, please refer to the accompanying guide for additional details.

Submission

- The assignment should comprise a concise user manual detailing the steps to execute the code for training, evaluation, and label generation. Additionally, it should feature a brief discussion covering implementation choices, potential variations across the explored models, and an analysis of performance across the assessed datasets. The document's length should not exceed three pages, with a font style of Calibri and a minimum font size of 11pt.

- The assignment will be done in groups of 2 or 3 people.

- The assignment's deadline for submission on the virtual campus is November 3, 2025, at 15:30h at the latest. Only one member of the group is required to submit the assignment. Submissions after the deadline will automatically receive a score of 0 out of 10 points.

- Defenses for the assignment will be scheduled during the following weeks after the submission, within the laboratory hours designated by the professor. All group members must be present during the defense; otherwise, the absent member(s) will receive a score of 0 out of 10 points for the assignment. If there are scheduling conflicts due to personal or professional obligations, please contact the professor in advance to arrange an alternative time.

- This assignment accounts for 20% of the overall course grade.