Statistical NLP — Homework Exercise 2

Jonas Belouadi, Philipp Cimiano
Semantic Computing Group
Bielefeld University
Winter Semester 2023/2024

Note on Submission You are allowed to submit in groups of two people. When you want to submit in a group put the name of your colleague in the comment field. Only one person needs to submit.

All solutions have to be uploaded together as a single zip file to LernraumPlus. Solve the exercises by completing the functions in file `exercise_sheet2.py`. Provide some information about how to execute your Python code. Non-code answers should be included in a PDF-file. **For this exercise, you are only allowed to import modules when explicitly stated in the task description.**

Task 1 [2.5+2.5 points]

In this exercise sheet we will solve the task of named-entity recognition by using HMMs. The file `corpus_ner.zip` contains a corpus annotated in the IOB schema. Each line in this file contains two columns separated by an ASCII space character, the first column contains the token and the second one the label of the token. Sentences are separated by empty lines. The set of states S contains all labels contained in the provided corpus and the set of emission symbols K contains all tokens contained in the corpus. Add the special token `<unknown>` for unknown tokens to the set of all tokens. Treat tokens occurring only once in the corpus as unknown, i.e., replace them by the token `<unknown>`. You may use the provided function `import_corpus` for importing the corpus.

a) Implement the Python functions `initial_state_probabilities`, `transition_probabilities` and `emission_probabilities` representing the parameters $P_\pi(\cdot)$, P_A and P_B, respectively, of an HMM. Each of these functions takes one argument (among others) named `internal_representation`. This argument is a data structure of an internal representation of the parameters $P_\pi(\cdot)$, P_A and P_B, respectively. It is up to you which data structure you use. For example, you can use a Python dictionary for representing the parameter $P_\pi(\cdot)$. We denote the data structures for the parameters $P_\pi(\cdot)$, P_A and P_B by D_π, D_A and D_B, respectively.

b) Next, implement the Python functions `estimate_initial_state_probabilities`, `estimate_transition_probabilities` and `estimate_emission_probabilities` for learning the parameters of the data structures D_π, D_A and D_B, respectively.

See `exercise_sheet2.py` for further information.

Task 2 [15 points]

Implement the Viterbi algorithm by defining the Python function `most_likely_state_sequence`. Instead of maximizing $P(X|O,\mu)$, maximize $\log P(X|O,\mu)$, i.e., replace multiplication by summation and probabilities by log probabilities (you are allowed to use the `log` function from the `math` library for that). Furthermore, replace each word of the input sentence not occurring in the corpus by the special word `<unknown>`. Make use of the Python functions you implemented in Task 1.