1 Viterbi algorithm for second-order HMM

In the lecture we saw the formulation of the Viterbi algorithm for a first-order Hidden Markov Model. In a second-order HMM transition probabilities depend on the two previous states, instead on just the single previous state, that is we use the following independence assumption:

\[ P(y_i|x_1, \ldots, x_{i-1}, y_1, \ldots, y_{i-1}) = P(y_i|y_{i-2}, y_{i-1}). \]

For this exercise you should formulate the Viterbi algorithm for a decoding a second-order HMM.

2 HMM POS-tagger

In this exercise you will implement a simple supervised HMM Part-of-Speech tagger. Your implementation should learn the model parameters from annotated data, and then decode unlabeled data with the learned model.

- Learning: estimate emission and transition parameters from sentences marked with POS-tags. For the purpose of this exercise it is enough to use a simple add-one method for smoothing out zero probabilities. In addition, in order to deal with unseen words, you can do the following: find words which have frequency less than 2 in the training set. Then replace occurrences of those words in both the training and test set with a special UNKNOWN token prior to estimating the probabilities.

- Decoding: implement the (first-order) Viterbi algorithm to find the most likely POS-tag sequence for a new sentence.

The training and test data are child-directed transcribed and POS-tagged sentences from the CHILDES corpus: \texttt{training set} and the \texttt{test set}.

Train your model, run it on the test data and report the per-token accuracy.

Please send your solutions by Tuesday Jan 26 to:

- gchrupala@lsv.uni-saarland.de if attending the Thursday 12:15 session.
- Munir.Georges@lsv.uni-saarland.de if attending the Friday 14:15 session

Important: Please use PDF as a document format. If you need to compress files, use ZIP or GZIP.