Prof. Dr. Andreas Hotho,
M.Sc. Janna Omeliyanenko
Lecture Chair X for Data Science, Universität Würzburg

2. Exercise for "Sprachverarbeitung und Text Mining"

12.11.2021

1 Knowledge Questions

1. What is Part of Speech Tagging and what may Part of Speech Tagging be used for?
2. What is the difference between open and closed classes of word types in POS-tagging?
3. Which two typical approaches are used for POS-tagging?
4. Which two types of probabilities are needed in the modeling approach of the Hidden Markov Model? What elements does a Hidden Markov Model generally consist of?
5. What does the Markov-assumption state when applied to POS-tagging?
6. Which POS-tagsets do you know for the German and English language?
7. All POS-tagging methods presented in the lecture require a list (dictionary) that lists all possible POS-tags for a word. However, such lists are usually not complete. Give a solution for POS-tagging with unknown words.
8. Name three types of evaluation of POS-tagging methods.

2 POS Tagging

Part-of-speech taggers model the words as observations (observed variables), and the associated word types (part-of-speech tags) as hidden variables (hidden states). We assume in the following a highly simplified set of generalized word types (tag-set):

DET	Determiner	the,a,...
N	Noun	year,home,costs,time,...
PRO	Pronoun	he,their,you
V	Verb	said,took,saw

Using the table below, calculate the most likely sequence of tags for the following sentence:
„`I saw the saw`“

1. Apply the Viterbi algorithm.
2. Which are the main factors influencing the complexity of the calculation?

Assume uniformly distributed start transition probabilities.

from\to	DET	N	PRO	V
DET	0.05	0.9	0.05	0
N	0.1	0.7	0.05	0.15
PRO	0.05	0.5	0.05	0.4
V	0.5	0.1	0.4	0

w\t	DET	N	PRO	V
I	0	0.0002	0.1	0
saw	0	0.0008	0	0.1
the	0.1	0	0	0