Assignment
Language Models and Part-of-Speech Tagging

Problem 1

Questions:

- What is a statistical language model?
- What is the formula for a bigram language model?
- What is part-of-speech tagging?
- What do the following POS tags stand for?

  AT, BEZ, IN, JJ, MD, NN, NNP, NNS, RB, TO, VB, VBD, VBG, VBN, VBZ, WDT
- What are the tags of the following parts of speech (or morphosyntactic "feature bundles")?

  article, the word "is", preposition, adjective, modal, singular or mass non-proper noun, singular proper noun, plural non-proper noun, adverb, the infinitive marker "to", verb (base form), verb (past tense), verb (present participle), verb (past participle), verb (3rd sng. present), wh-determiner
- What are the two sources of information used in statistical POS tagging?
- What is the form of the word bias (emission) parameter in POS tagging?
- What is the form of the context (transition) parameter in POS tagging?
- What is the advantage of Laplace estimation compared to ML estimation?

Problem 2

Give examples for the following.

- Using word bias (emission) probability is useful for POS tagging.
- Using context (transition) probability is useful for POS tagging.

Problem 3

Estimate the word bias (emission) parameters $P(\text{authorization}|\text{NN})$ and $P(\text{restrict}|\text{VB})$ based on the following training text. Give ML und Laplace estimates.

> The/DT bill/NN intends/VBZ to/TO restrict/VB the/DT RTC/NNP to/IN Treasury/NNP borrowings/NNS only/RB ,/, unless/IN the/DT agency/NN receives/VBZ specific/JJ congressional/JJ authorization/NN ./.

Problem 4

Estimate the context (transition) parameters $P(\text{VB}|\text{TO})$ and $P(\text{TO}|\text{VB})$ based on the following training text. Give ML und Laplace estimates.

> The/DT bill/NN intends/VBZ to/TO restrict/VB the/DT RTC/NNP to/IN Treasury/NNP borrowings/NNS only/RB ,/, unless/IN the/DT agency/NN receives/VBZ specific/JJ congressional/JJ authorization/NN ./.

Problem 5

Estimate the probabilities $P(\text{wenige}|\text{nur})$ and $P(\text{in}|\text{nur})$ based on the following training text. Give ML und Laplace estimates.

> nur wenige Zoos halten Greifstachler , in Deutschland nur der Frankfurter .

Problem 6

Suppose a speech recognition program returns two recognition hypotheses h_1 and h_2 for a spoken sentence.

- h_1: ich komme vom Hauptbahnhof
- h_2: ich komme vom Haupt Bahn Hof

A language model P_{LM} trained on a large German corpus will assign a much higher probability to h_1 than to h_2:

$$P_{\text{LM}}(h_1) \gg P_{\text{LM}}(h_2)$$

Explain why.