Universität Bielefeld Semantic Computing Group

Exercise Sheet

Statistical Natural Language Processing

Exercise Sheet 1

Due Date: October 31, 10 pm

Note on Submission

All solutions have to be uploaded together as a single zip file to LernraumPlus. Provide some information about how to execute your Python code.

Exercise 1 [3+6+1 points]

Download the file `corpus.zip` from LernraumPlus. The file `corpus.txt` contains the corpus we will use in this worksheet, there is exactly one sentence in each line of this file. A sentence is a sequence $w_1, \ldots, w_N$ of words, where w_1 is the first word in the sentence, w_N is the last word and N is the number of words in the sentence. Now we define some distributions of words for the provided corpus:

- $P(w)$ is the distribution of all words in the corpus
- $P(w_i|w_{i-1})$ is the distribution of words given the previous word in a word sequence is w_{i-1}
- $P(w_i|w_{i-1}, w_{i-2})$ is the distribution of words at position i in a word sequence given the word at position $i-1$ is w_{i-1} and the word at position $i-2$ is w_{i-2}

a) First of all you need to preprocess the corpus. Therefore implement a Python function which takes a single string as input (representing a sentence) and returns a sequence of words. You may ignore commas, semicolons and colons. Do not use any NLP related libraries!

b) Provide Python code for representing and learning the distributions $P(w)$, $P(w_i|w_{i-1})$ and $P(w_i|w_{i-1}, w_{i-2})$.

c) How does the number of parameters of these distributions scale with the number of different words in the corpus? Explain your answer!

Hint: Introduce special words to model the beginning and the end of a sentence!

P. Cimiano · C. Witte

Exercise 2 [3+6+1 points]

a) Implement Python functions for drawing samples from the distributions $P(w)$, $P(w_i|w_{i-1})$ and $P(w_i|w_{i-1}, w_{i-2})$ according to the algorithm presented in the lecture. Make use of your solution of Exercise 1.

b) Use the statistical information of the provided corpus to implement three different sentence generators, i.e., use the distributions $P(w)$, $P(w_i|w_{i-1})$, $P(w_i|w_{i-1}, w_{i-2})$ and your code of a) to generate single words. In other words, your first sentence generator should use the distribution $P(w)$, the second one $P(w_i|w_{i-1})$ and the third one $P(w_i|w_{i-1}, w_{i-2})$.

c) Describe the results of the three sentence generators you implemented. Try to explain the results.