ELEC-E5550 Statistical Natural Language Processing, exam 10.4.2019

- *There are 5 questions, each is worth the maximum of 6 points, so total 30 points*
- *You have 3 hours to complete the exam.*
- *You may use a scientific calculator.*
- *No additional material is allowed.*

Question 1

Explain the following terms (1p each)

a) Zipf's law

b) Term frequency – inverted document frequency (tf.idf) weighting

c) Minimum description length (MDL) principle

d) Perplexity value

e) Beam search

f) Probabilistic context free grammar (PCFG)

Question 2

a) Describe a method for Part-of-Speech tagging. (2p)

b) Describe how to solve ambiguous words in this method. (2p)

c) What are the advantages and the disadvantages of this method compared to other methods. (2p)

Question 3

Calculate the 4-gram BLEU score for the two translation hypotheses SYS1 and SYS2, given the reference translation REF.

REF : *it cannot serve as a basis for the establishment of a european constitution*

SYS1: *she can be used as a basis for the installation of a european constitution*

SYS2: *it cannot into a basis for the european constitution*

Question 4

What are the essential components of a large-vocabulary continuous speech recognition system? Describe shortly what kind of methods are typically used in each component. You should identify at least five components.

Question 5

Neural Network Language Models take a word sequence as input and output probabilities over a vocabulary. Words are popularly represented as one-hot vectors i.e. vectors with all zeros except a one at the specified word index.

1. What are the disadvantages of using one-hot representation? (1 point)

2. Suggest two alternatives that you can apply to overcome these disadvantages? (2 point)

Neural Network Language Models include e.g. Feedforward Neural Network (FFNN) and Recurrent Neural Network (RNN).

3. What are the major architectural differences between FFNN and RNN? (1 points)

4. What are the advantages and the disadvantages of using RNN instead of FFNN? (2 point)