ELEC-E5550 Statistical Natural Language Processing, exam 5.4.2017

- *There are 5 questions, each question is worth the maximum of 6 points, 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

There are 50 000 documents in a database. A user makes a query that should give five relevant documents. Two competing search engines return ordered lists of ten documents. The relevant ones are marked with **R**, and non-relevant ones are marked with N.

Engine 1: 1**R**, 2**R**, 3N, 4**R**, 5N, 6N, 7N, 8N, 9N, 10N

Engine 2: 1N, 2N, 3**R**, 4**R**, 5N, 6N, 7N, 8**R**, 9**R**, 10**R**

Calculate the following evaluation measures for both engines (1p each):

a) Precision

b) Recall

c) Accuracy

d) Error

f) F-measure (balanced, α=0.5)

e) Uninterpolated average precision

Question 3

What is Sentiment Polarity Detection and in what kind of applications can it be used?

What are the essential steps in training a modern Sentiment Polarity Detection system?

Describe shortly what kind of methods are typically used in these steps.

You should identify at least 6 steps.

Question 4

Use the Viterbi algorithm to calculate the most probable phoneme sequence for the observations given in Table 1. The observations in this exercise are given in the form of emission probabilities. In reality they would be estimated for features representing the observations using a probabilistic classifier such as GMM. The transition probabilities of the HMM are given in Table 2.

Table 1. Observation / emission probabilities.

time model	t0	t1	t2	t3
_	0,4	0	0	0
/d/	0,5	1	0	0
/g/	0,1	0	0	1
/i/	0	0	0,5	0
/o/	0	0	0,5	0

Table 2. Transition probabilities. Each row indicates the preceding HMM (first) and each column indicates the following HMM (second).

second first	_	/d/	/g/	/i/	/o/
START	0,5	0,2	0	0,12	0,18
_	0,8	0,1	0,025	0,05	0,025
/d/	0,05	0,8	0	0,05	0,1
/g/	0	0	0,8	0,1	0,1
/I/	0,1	0	0,2	0,7	0
/o/	0,05	0	0,25	0	0,7

Question 5

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