ELEC-E5550 Statistical Natural Language Processing, exam 4.4.2018

- *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

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

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

c) Describe how to compute a probability of a given tag sequence for a sentence. (2p)

Question 2

Consider the following simple language consisting of a set of English words, with the corresponding observed frequencies in corpus C:

{ any (3), thing (2), some (2), something (3), anyone (2), anything (2) }.

Two alternative morph lexicons are introduced for the language. Both are encoded using characters from a standard alphabet (26 letters + space) as follows,

Lexicon 1: any_thing_some_one__ (4 morphs, total length 20 characters)

P(Lexicon 1) = (1/27)^20 = 2.4e-29

Lexicon 2: any_thing_some_something_anyone_anything__ (6 morphs, total length 42 characters)

P(Lexicon 2) = (1/27)^42 = 7.6e-61

Spaces are marked with underscore for clarity. A single space indicates a morph boundary. Two spaces indicate the end of the lexicon.

a) For both lexicons, estimate the probability of the corpus C given the lexicons, ie. P(C|Lexicon 1) and P(C|Lexicon 2). Assume that the distribution of the morphs in the lexicon is uniform, ie. P(morph|Lexicon L) = 1/|Lexicon L| and a word break morph is included in the vocabulary to denote word breaks. Which lexicon gives a higher probability for the language? You can approximate the probabilities, but show your computations. (3p)

b) Estimate the Maximum a Posteriori (MAP) probability for the lexicons 1 and 2, ie. P(Lexicon 1| C) and P(Lexicon 2|C). Which of the lexicons is a more likely model of the language? You can approximate the probabilities, but show your computations. (3p)

Question 3

a) What is a vector space model and in what kind of applications it can be used? (2p)

b) Describe the main steps in building a vector space model. You should identify at least 6 steps. Explain two methods which can be successfully used in each step. (4p)

Question 4

a) Use Good-Turing smoothing to estimate the probability of catching next any fish species you have not caught yet, if you have already got 3 perches, 1 pike and 1 zander. Show your computations. (2p)

b) Explain why smoothing is important for n-gram language models. (2p)

c) Describe the principles (including equations) of Good-Turing smoothing and discuss its strengths and weaknesses for estimating the probability of a word n-gram. (2p)

Question 5

What are the essential steps in training a phrase-based statistical machine translation system (e.g. Moses)? Describe shortly what kind of methods are typically used in these steps. Assume that the input data is a paragraph-aligned parallel corpus. You should identify at least six steps.