T-61.5020 Statistical Natural Language Processing, exam 19.9.2016

- *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.*

$$P(C|Lex1) = \frac{P(Lex1|C)P(C)}{P(Lex1)}$$

Question 1

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 2

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 2 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)

(0 0 •) •00

Question 3

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 $tp = 3$ $fp = 7$ $fn = 2$

Engine 2: 1N, 2N, 3**R**, 4**R**, 5N, 6N, 7N, 8**R**, 9**R**, 10**R** $tp = 5$ $fp = 5$ $fn = 0$

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

a) Precision

b) Recall

c) Accuracy

50000 - 10

- fn = 2

d) Error

f) F-measure (balanced, α=0.5)

e) Uninterpolated average precision

Question 4

Popular sequence tagging methods include the generatively trained Hidden Markov Models (HMMs) and the discriminatively trained Conditional Random Fields (CRFs).

a) Both HMMs and CRFs are graphical models. Compare their corresponding graphs. How are they similar, what are their differences, and how does information flow in the graph? (2p)

b) Compare their sensitivity to noisy features. Motivate your answer. (2p)

c) Compare their estimation effort. Motivate your answer. (2p)

No equations are required.

Question 5

What are the essential steps in training a modern 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.

selecting data
training
sentence alignment
word alignment
phrase alignment
evaluation

distortion
fertility
reordering