T-61.5020 Statistical Natural Language Processing, exam 25.4.2015

- *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

What is a vector space model and in what kind of applications it can be used? Describe the main steps in building a vector space model and explain two methods which can be successfully used in each step.

Question 2

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 3

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.

Question 4

Each cell in Table 1 shows bigram counts of seven words in a text corpus. Table 2 shows the unigram counts for the same words. Calculate bigram probabilities for the following word pairs. Show your calculations.

A: "I want"

B: "eat Chinese"

C "Chinese lunch"

Solve bigram probabilities for following word pairs that do not have any bigram examples in the corpus. You can choose which one of these two methods you apply: 1. back off to unigram, 2. add-one smoothing. If you use back-off you can use back-off weight b_w=0.1 . Total number of word types in the vocabulary is 2000 and total number of words in the whole text corpus is 500 000. Show your calculations.

A: "Chinese eat"

B: "lunch lunch"

C: "I Chinese"

Table 1. Bigram counts modified from the Jurafsky & Martin example. Preceding context word is given in the first column and the current word in the first row.

	I	want	to	eat	Chinese	food	lunch
I	9	1072	0	13	0	0	0
want	4	0	780	0	5	9	5
to	3	0	11	855	3	0	14
eat	0	0	2	0	19	2	57
Chinese	2	0	0	0	0	118	1
food	17	0	15	0	0	0	0
lunch	6	0	0	0	0	1	0

Table 2. Unigram counts modified from the Jurafsky & Martin example.

I	3442
want	1212
to	3123
eat	920
Chinese	199
food	1405
lunch	450

Question 5

Explain what are kernel methods and how they can be used to analyze text documents. Give two concrete examples of kernels suitable for text classification. Discuss the capabilities and limitations of them.