UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Exam in INF5830 - Natural language processing
Day of exam: 16 December 2011
Exam hours: at 09:00 – 4 hours
This examination paper consists of 4 pages including this.
Appendices: Statistical table – 3 pages
Permitted materials: None

Make sure that your copy of this examination paper
is complete before answering.

- **You may answer in English, Norwegian, Danish or Swedish.**
- **You should answer all questions. The weight of the various questions are indicated.**
- **You should read through the whole set to see whether anything is unclear so that you can ask your questions to the teachers when they arrive.**
- **If you think some assumptions are missing, make your own and explain them!**

1 Evaluation (10%)

Lee is experimenting with word sense disambiguation. He uses the *hard*-data from the famous *hard-line-serve*-corpus, and reserves 433 (1 out of 10) occurrences for testing and uses the rest for training. This is the result from one of his runs.

	Correct class		
	Sense1	Sense2	Sense3
Assigned Sense1	341	28	23
Assigned Sense2	6	15	0
Assigned Sense3	2	1	17

For each occurrence of *hard*, Lee is interested in whether it is classified correctly or not. What is the accuracy of the classifier in this sense (i.e., micro-averaged accuracy).

Suppose we are only interested in how well the classifier is doing as a classifier for Sense3. This corresponds to merging Sense1 and Sense2. What is the accuracy, recall and precision for the classifier for Sense3? You may present the answers as fractions.

2 Classification and clustering (30%)

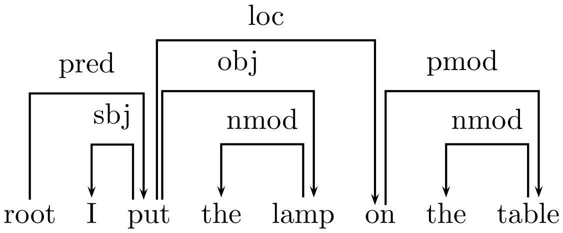
(a) Explain in a few sentences the difference between supervised and unsupervised learning.

(b) Explain how Rocchio classification works. What are the strengths and weaknesses of the Rocchio classifier?

(c) Explain the K-means clustering algorithm.

3 Dependency syntax and parsing (40%)

(a) Consider the dependency graph for the English sentence *I put the lamp on the table*:



Discuss the notion of **syntactic head** and describe at least three different criteria for head status, illustrating with examples taken from the dependency representation above.

(b) Does the dependency graph above satisfy the constraint of **projectivity**? Why/why not? Can you provide an example of a syntactic construction that requires a non-projective representation?

(c) Nivre's arc eager algorithm operates with four parse **transitions**, two of which are parameterized by the dependency relation r: Shift, Reduce, Left-Arc$_r$ and Right-Arc$_r$.

Show the transition sequence used to derive the dependency graph above, at each step providing the transition employed (Shift, Reduce, Left-Arc$_r$, Right-Arc$_r$), as well as the contents of the stack and queue. How does the algorithm ensure projectivity?

(d) What are **proto-roles** in the sense of Dowty (1991) and how do these account for the mapping between arguments and syntactic functions for our example sentence *I put the lamp on the table*?

4 Experiments (20%)

Kim has constructed a system for classifying textual entailment. She has read that the best classifiers have 0.72 accuracy when tested on test material which is evenly split between entailment and non-entailment examples, and this is the baseline she wants to beat. Since Kim is very careful and avoids seeing the test items herself, she has got her fellow students to construct a test base of 50 examples of entailment and 50 examples of non-entailment, annotate the examples and shuffle them. We may assume that these test examples can be considered a random sample. When she runs her classifier on the test set, it classifies 75 out of the 100 items correctly. At first Kim gets very excited and wants to publish her results immediately. But then she remembers vaguely something about statistical significance and comes to you for help.

(a) Is Kim's result statistically significantly better than the baseline at the 0.05 level? Since we do not use computers at this exam you may use 0.2 for $0.72 \times (1 - 0.72)$ (instead of 0.2016). You may also use 0.45 for $\sqrt{0.72(1 - 0.72)}$ (instead of 0.4490).

(b) Kim is a very careful student and thinks the 0.05 level is too little. She wants a result which is statistically significant at the 0.01 level. She realizes that she will need a larger test set for this and that she must bribe her fellow students into marking up some more examples. But approximately how many test items are necessary to show that a raise in performance from 0.72 to 0.75 (i.e. 3 per cent points) is statistically significant at the 0.01 level?

END