**ReMa L&C, Introduction to Language and Speech Technology**

**Final Exam, Jan 10, 2018**

This exam consists of 3 questions. Every question is worth 10 points. Subquestions are marked for their points.

First read all questions carefully and then choose in which order you will answer them. You can put the answers to the questions (and even parts of questions) in any order you want, as long as you clearly mark which answers belong to which questions. Also put your name and student number on any sheet of paper and/or computer file you hand in.

Please note that answers should be supported by argumentation, and where useful you should include examples. However, restrict yourself to relevant arguments and do not just try to write down everything you know. The presence of excessive irrelevant information may lead to a deduction of points.

**Background information for the questions: Biomedical text.**

Most of the NLP techniques described in the book are explained with examples taken from rather everyday language use, or from simple specific domains, and some techniques may well be limited to those types of language. We can therefore ask ourselves what happens if we try to process specialized and in places very complex text. In this exam, we will take a look at biomedical language use, as exemplified in the paper

http://www.sciencedirect.com/science/article/pii/S0002929707637860

which I brought to your attention last Wednesday. In each question, we will focus on a different level of detail.

1) Application: Machine Translation

   a. Discuss the relative strengths and weaknesses of direct, transfer, and interlingual machine translation (25.2) for biomedical text. (3 points)
   b. Describe how you yourself would set up a “classical”, i.e. non-statistical, MT system (25.2) targeted at biomedical text. Explain your choices. (3 points)
   c. Describe how well you expect statistical MT to work on biomedical text. Explain your expectations. (2 points)
   d. Describe how you would adapt an existing statistical MT system, aiming for fully automatic high quality translation of biomedical text. (2 points)

2) Component tasks: Information Extraction. In the book, various techniques are proposed for a number of tasks which can be included in an information extraction pipeline. Discuss for each of the tasks listed below how you would adapt the techniques described in the book for application in the biomedical domain. If you think no adaptation is needed, explain why not and how then you would apply (a selection of) the standard techniques.

    a. Word sense disambiguation (20.1-20.5; 2 points)
    b. Semantic role labeling (20.9; 2 points)
    c. Relation detection (22.2; 3 points)
    d. Event detection (22.3; 3 points)

3) Algorithms: Alias detection. An important feature for coreference resolution represents the fact that two strings are each other's alias, i.e. that both are names referring to the same entity (21.7). The paper mentioned in the introduction above shows several forms of aliases that can be found in the biomedical domain.

    a. Discuss the usefulness of the Levenshtein algorithm for alias detection in the biomedical domain. (2 points)
    b. A possible form of an alias is the acronym, as exemplified by *COMT*. It might be possible to use an IOB-tagging approach to identify the acronym characters inside the full name. Explain how you could use the Viterbi algorithm for such IOB-tagging. (2 points)
    c. Such a Viterbi approach could be made much more effective if we extend the HMM to an MEMM, and allow additional features to determine the probabilities at each position. Discuss the additional features you would use in this case. (3 points)
    d. Aliases might also be discovered on the basis of similarities of context. Discuss the usefulness of context vectors for alias detection in the biomedical domain. (3 points)