NTNU – Trondheim
Norwegian University of
Science and Technology

Department of Computer Science (IDI)

Examination paper for TDT4310: Intelligent Text Analytics and Language Understanding

Academic contact during examination:	**Prof. Björn Gambäck**
Phone:	**+46 70 568 15 35**
Examination date:	**May 23rd, 2018**
Examination time (from-to):	**09:00 – 13:00**
Permitted examination support material:	**D (no books or notes allowed)**

Other information: Since all the course material and the lectures were in English, the exam is too. However, answers may of course be given in either English or Norwegian.

Read the questions carefully. If you consider that some information needed to solve a question is missing, give a short account of the assumptions you have found it necessary to make.

The maximum number of points on the exam: 80. The exam counts for 80% of the overall course grade. There are 20 questions, each counting for 4 points. However, this does neither imply that the questions are at an equal level of difficulty, nor that the answers to the questions should be of about equal length. The opposite is rather the case, so some questions are most certainly easier than the others and require shorter answers. It will thus make sense to first reply to those. Note though that the questions are not ordered by level of difficulty, so read all questions before starting to answer the ones you consider “low-hanging fruit”.

Language:	**English**
Number of pages (front page excluded):	**1**
Number of pages enclosed:	**2**

Checked by:

Date Signature

Students will find the examination results in Studentweb. Please contact the department if you have questions about your results. The Examinations Office will not be able to answer this.

TDT4310: Exam Questions 2018 [maximum points: 80]

1) **Give some reasons for processing human languages.**

2) **Suppose you wanted to create an artificial human language (such as Esperanto and Interlingua) that could be used for communication by everybody on the planet Earth regardless of mother tongue. Which would be most important issues to consider?**

3) **What is the Distributional Hypothesis?**

4) **Describe Frege's Principle.**

5) **What is Lambda Calculus (and lambda reduction)?**

6) **What is "Grounding"? How does it relate to the "Principle of closure" (Clark)?**

7) **When evaluating NLP systems, we use precision, recall and F-score. What are those?**

8) **What is cross-validation and how is it used?**

9) **Discuss the influence of the size of the available data on the choice of text classification methods.**

10) **A crucial factor for the coverage and performance of a part-of-speech tagger is its ability to handle words unknown to the system. Suggest two different methods that make it possible to assign tags even to those words.**

11) **What are the basic components of an opinion?**

12) **Briefly describe how you could build a tool to determine if a movie review is positive or negative.**

13) **Statistical Machine Translation systems basically consist of two language models. Briefly describe the difference between them and how they relate to Bayes' Theorem.**

14) **What is the noisy channel model? Describe how it is utilized in statistical machine translation.**

15) **The term FAHQGPMT came about partially as a parody of all the abbreviations in the computer society. It does, however, introduce three important constraints on a machine translation system. Select *one* of them and briefly describe how relaxing it can make MT feasible today.**

16) **Describe the differences between a machine translation system using an interlingua approach and a system using a transfer-based approach. Discuss some advantages and problems with each strategy.**

17) **Describe the bag-of-words model and n-gram language models. How do they relate to each other?**

18) **Why is the independence assumption useful for language models?**

19) **A common statement concerning natural language grammars is that "all grammars leak".**
A related statement is "natural languages are inherently ambiguous".
Discuss different techniques to suppress ungrammatical interpretations in a grammar, and techniques to give ambiguous sentences different interpretations.

20) **The company *Skrivefeil AS* wants you to build a spelling correction system for Norwegian using a lexicon-based strategy, i.e., an approach that assumes that all words can be found in the lexicon. Why is this not a good idea? Briefly outline a better strategy.**

Students will find the examination results in Studentweb. Please contact the department if you have questions about your results. The Examinations Office will not be able to answer this.