Exam in
TDT4275 – Natural Language Interfaces
(Naturlig språklig grensesnitt)

Saturday 9. June 2012, 09:00 – 12:00

The exam was created by Prof. Björn Gambäck and quality assured by Erwin Marsi.
Contact person during the exam: Björn Gambäck (office: 735 93354, mob: +46 70 568 1535)

Language: English (answers may also be given in Norwegian)

No books or notes are allowed.
Tillatte hjelpemidler: D
Ingen trykte eller håndskrevne hjelpemidler tillatt.


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

Les oppgaveteksten nøye. Finn ut hva det spørres etter i hver oppgave.

Dersom du mener at opplysninger mangler i en oppgaveformulering gjør kort rede for de antagelser og forutsetninger som du finner det nødvendig å gjøre.
The exam consists of ten questions. Each question gives as many points as its number. Thus, the maximum number of points is 55.

No books or notes are allowed. Since all the course material and the lectures were in English, the exam is too. However, you may of course answer in either English or Norwegian.

1. What are the more prominent word classes that carry sentiment? [1p]

2. Briefly describe the two concepts precision and recall. [2p]

3. What some would claim to be the most fundamental linguistic event of all time has been described as follows:

   “The Lord came down to see the city and the tower, which the children of men builded. And the Lord said, Behold, the people is one, and they have all one language; and this they begin to do; and now nothing will be restrained from them, which they have imagined to do. Go to, let us go down, and there confound their language, that they may not understand one another’s speech.”

   (Illustrated by Gustave Doré in the painting “The Confusion of Tongues”.)

Supposedly, 72 different languages (and 72 different tribes / cultures) were created at this event. If you were to confound one language into 72, which parameters would you play with? Suggest at least five binary valued parameters and one parameter with three (or more) sensible values. [3p]
4. According to the *Lebor Gabála Érenn*, the king Fénius Farsaid took the best pieces of all the 72 languages and created Goídelc, the Gaelic language.
   a) Which language is the best is a subjective issue, but suppose that Fénius had been a Computational Linguist and suggest three basic features which he should have included in the language (i.e., properties which would make a language “good” from a Computational Linguistic perspective). [3p]
   b) Fénius Farsaid did not lie down and rest, but went on to discover alphabets. First Hebrew, Greek and Latin, and finally Ogham (Old Irish) which supposedly is the most perfected - because it was the last. Fénius probably had not been told about Unicode and UTF-8, but suppose that you wanted to create the perfect writing system, from a Computational Linguistic perspective. Suggest at least one property that this writing system should have. [1p]

5. Before things got mixed up (i.e., the languages got confused), there supposedly was only one language, common to all.
   a) What would such a language be called if used in a machine translation setting? [1p]
   b) If Fénius would have wanted to create such a language, which problems would he have run into? [2p]
   c) Take the Vauquois triangle as basis for briefly describing two other machine translation approaches. [2p]

6. Bayes’ decision rule has turned up in many shapes during the course.
   a) Thus a statistical machine translation system consists of two models which can be derived from Bayes’ rule. Describe them and their relationship. [3p]
   b) In a similar vein, a Bayes’ rule-based speech recognition system relies on two models. Describe them and their relationship. [3p]

7. What does it mean that a dialogue system allows for *mixed initiative*? Explain how such a dialogue treatment could be achieved. [7p]

8. The company Predictor Ltd has released program for word prediction based on trigrams. The idea was that the program would suggest the next word based on the two previous words typed by the user. However, when the program was evaluated it soon became apparent that the probabilities for several plausible trigrams were zero. What was the reason for this? And why is this a problem? Suggest a method for solving the problem. [8p]
9. The University of Tromsø is well-known for aggressively trying to recruit NTNU students even before they have graduated. Hence you have been employed in their project to create a system for translating between Norwegian Bokmål and Northern Sami, after having made a wild claim to know the language. Describe how you would go about creating the machine translation system, supposing that you have no prior Sami resources available and that the only thing you actually know about the Northern Sami language is its orthography, as shown below.

<table>
<thead>
<tr>
<th>Aa Áá BbCc Čč Dd Đđ Ee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ff Gg Hh Ii Jj Kk Ll Mm</td>
</tr>
<tr>
<td>Nn Ññ Oo Pp Rr Ss Šš Tt</td>
</tr>
<tr>
<td>Ţţ Uu Vv Zz Žž</td>
</tr>
</tbody>
</table>

10. Electronic patient records contain a lot of interesting information; however, the access to the information is in general quite restricted and the hospitals rarely utilize the patient records to their full potential. Commonly, a patient record would contain a description (in words) of the patients overall health condition, followed by the diagnosis reached by the doctor (or “failure to diagnose” in case the patient has some symptoms that have gone undiagnosed). How would you go about helping a hospital create applications that could use the information stored in the patient records? Which sets of data would you create, and how? What types of representations would you use? In particular consider the following two use cases:

a) Suppose a doctor fails to diagnose and hence wants to see patient records with health condition descriptions similar to the current patient’s. How would you retrieve patient records with similar health condition descriptions? [5p]

b) Suppose the hospital wants to completely replace the doctors with a system, by suggesting diagnoses for new patients only based on their health condition descriptions. How would you go about doing that? Would you be able to claim that your system was completely reliable? (Why or why not?) [5p]

Good luck!

Björn