UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Exam in **INF5820 Language technological applications**
Day of exam: **7 June**
Exam hours: **0900-1300**
This examination paper consists of 4 pages.
Appendices: **0**
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 weights 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 Machine translation (20%)

In the literature, machine translation (MT) systems are traditionally sorted into three classes called, *direct, interlingua and transfer-based* MT, resp. Describe the three different strategies. Compare the three strategies to each other. What do you see as their main strengths and weaknesses?

2 Decoding (30%)

Say we are to translate between German and English. We are using phrase-based statistical MT. The task is to translate sentence (1) to sentence (2) based on the phrase correspondences shown in figure 1.

1. *er geht ja nicht nach hause*
2. *he does not go home*

There are also other candidate phrases; some of them shown in figure 2.

(a) To get the best translation we base our choice on combining several probabilities. What are the probabilities involved?

(b) Describe the main structure of the algorithm used for finding the most probable English sentence, i.e., the decoding algorithm.

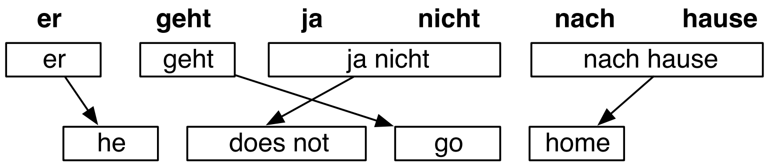

Figure 1:

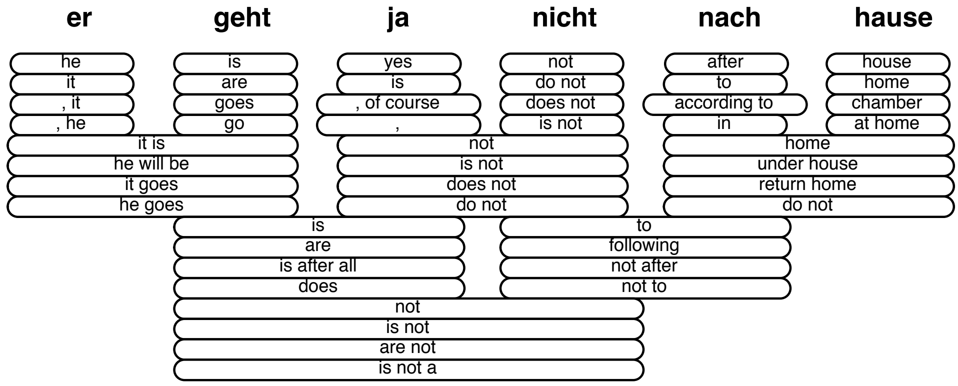

Figure 2:

(c) An exhaustive search for the best translation would be intractable. We therefore use various techniques for reducing the search space, including *recombination* and *pruning*. Explain how these work.

3 Spoken language (10%)

We have seen in the course that spoken language can have a markedly different structure than written language. Please illustrate this difference with at least two linguistic phenomena that are characteristic of spoken language. Provide a few sentences of explanation for each phenomenon that you mention.

4 Probabilistic modelling (20%)

You want to build a classifier that is able to classify utterances in two classes of dialogue acts: questions and statements. In order to do so, you want to use two complementary sources of information: word order (e.g. presence of inversion) and pitch. The relation between these two linguistic traits and the underlying dialogue act is however non-deterministic, so you decide to model the problem in a probabilistic manner, based on collected data.

You collect a total of 100 dialogue acts, and you manually annotate each dialogue act with two features:

- word order, with two possible values: *default* or *inverted*;
- ending pitch, with three possible values: *low*, *medium* or *high*.

The result of your annotation is summarised in the following table:

Type of dialogue act	Total count (on 100)	Word order		Pitch		
		default	*inverted*	*low*	*medium*	*high*
Questions	**20**	10	10	2	3	15
Statements	**80**	80	0	50	25	5

Given the information in the table, please derive the following probabilities:

- The prior probability $P(\text{dialogue act} = \textit{question})$ of a question without knowing any other information.
- The probability $P(\text{dialogue act} = \textit{question} | \text{word order} = \textit{default})$ of a question if we know that the word order is the default one.
- The probability $P(\text{dialogue act} = \textit{question} | \text{pitch} = \textit{high})$ of a question if we know that the ending pitch is high.
- And finally, the probability

  $$P(\text{dialogue act} = \textit{question} | \text{word order} = \textit{default}, \text{pitch} = \textit{high})$$

  of a question if we know that both the word order is default and the ending pitch is high. You can assume that the two features are conditionally independent of one another (i.e. naive Bayes assumption).

Please detail each reasoning step in your calculations.

5 Dialogue management (10%)

During the lecture on dialogue management, we went through various approaches that have been proposed to formalise the dialogue management problem. Some approaches are "hand-crafted" (they require a system developer to manually encode a set of rules or system flow) while other approaches rely on machine learning to automatically learn the best action to perform from data. What are the respective advantages and disadvantages of these two types of approaches? Please provide at least one advantage and one disadvantage for each type.

6 Speech synthesis (10%)

What is the *text analysis* step in speech synthesis? Explain in a few sentences what are the inputs and outputs of this step, and how it works.

END