ℹ Frontpage

IN4080 Natural Language Processing

Fall 2021

Monday, December 13
15:00 AM - 19:00 PM (4 hours)

All questions should be answered!
Each question is assigned a weight which is indicated.
The maximum number of points for the whole set is 100 points.

Permitted materials: None
An on-screen calculator is available.

You may answer in English, Norwegian, Danish or Swedish.

1(a) Tokenization

One of the first steps in text processing is (word) tokenization. What is tokenization? With the sentence (*) from the Brown corpus as example, discuss the decisions a tokenizer have to take, and where different tokenizers may differ.

(*) "You shouldn't smoke so much '' , he said , unconsciously imitating Victoria's holier-than-thou voice.

Fill in your answer here

Maximum marks: 5

1(b) Lemmatization

b) What is a lexeme? What is meant by lemmatization in NLP? What will (*) look like after lemmatization?

(*) "You shouldn't smoke so much '' , he said , unconsciously imitating Victoria's holier-than-thou voice.

Fill in your answer here

Maximum marks: 10

2 Semantic relations

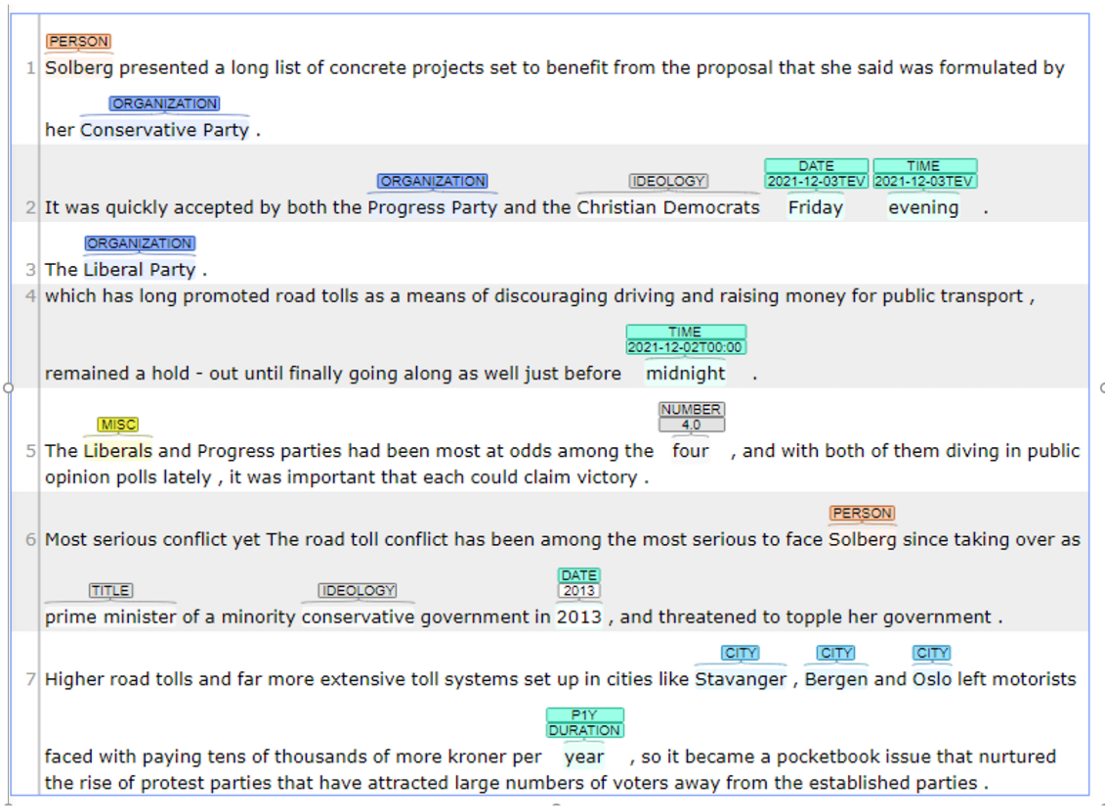
a) Which sematic relations are there between the words in each of the following pairs?

a. Horse – animal
b. Horse – cow
c. Horse – saddle
d. Big – small
e. Big – large

Fill in your answer here

Maximum marks: 5

3 Output from the Stanford corenlp



1 Solberg presented a long list of concrete projects set to benefit from the proposal that she said was formulated by her Conservative Party .

2 It was quickly accepted by both the Progress Party and the Christian Democrats Friday evening .

3 The Liberal Party .

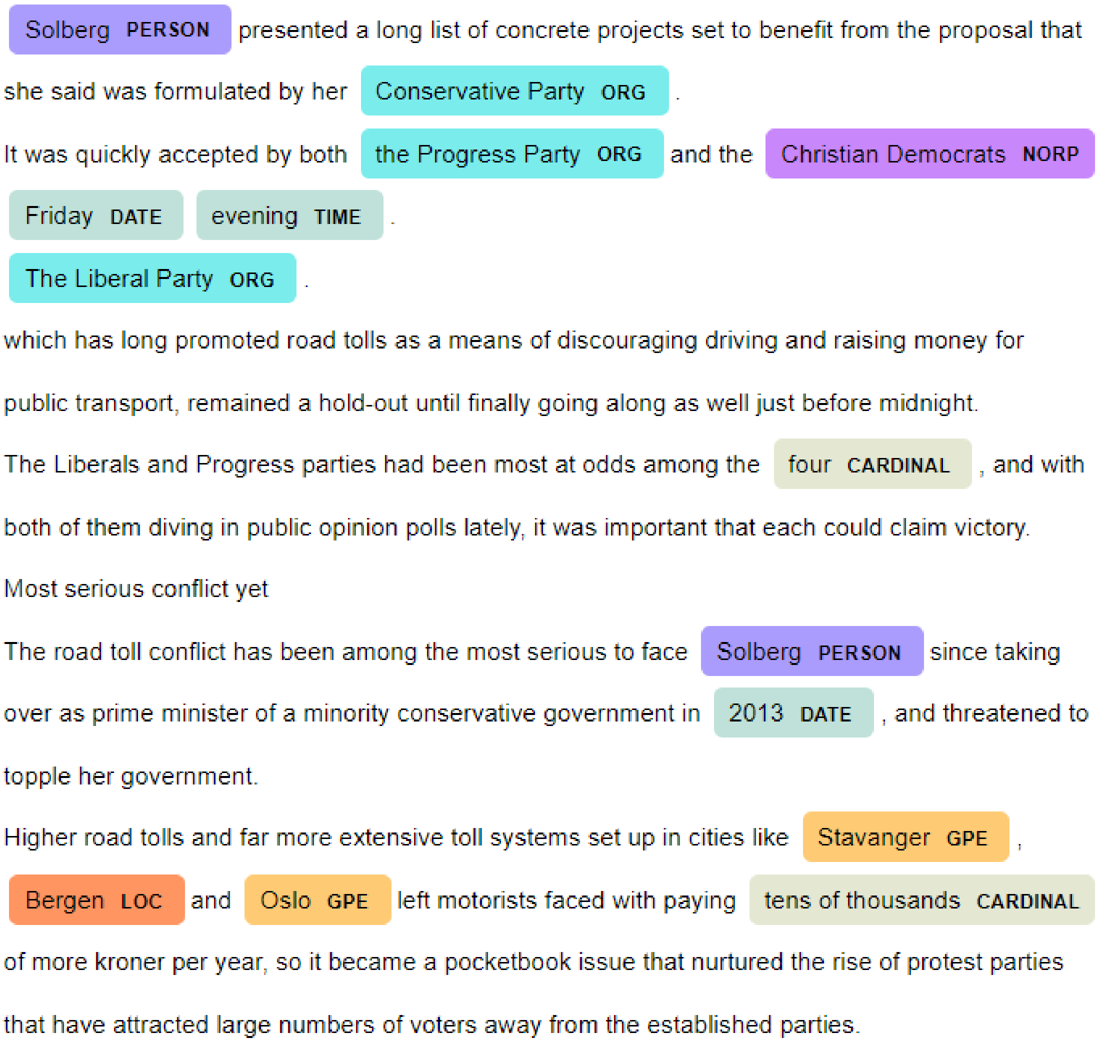
4 which has long promoted road tolls as a means of discouraging driving and raising money for public transport , remained a hold - out until finally going along as well just before midnight .

5 The Liberals and Progress parties had been most at odds among the four , and with both of them diving in public opinion polls lately , it was important that each could claim victory .

6 Most serious conflict yet The road toll conflict has been among the most serious to face Solberg since taking over as prime minister of a minority conservative government in 2013 , and threatened to topple her government .

7 Higher road tolls and far more extensive toll systems set up in cities like Stavanger , Bergen and Oslo left motorists faced with paying tens of thousands of more kroner per year , so it became a pocketbook issue that nurtured the rise of protest parties that have attracted large numbers of voters away from the established parties .

Output from Spacy:



Solberg PERSON presented a long list of concrete projects set to benefit from the proposal that she said was formulated by her Conservative Party ORG .

It was quickly accepted by both the Progress Party ORG and the Christian Democrats NORP
Friday DATE evening TIME .

The Liberal Party ORG .

which has long promoted road tolls as a means of discouraging driving and raising money for public transport, remained a hold-out until finally going along as well just before midnight.

The Liberals and Progress parties had been most at odds among the four CARDINAL , and with both of them diving in public opinion polls lately, it was important that each could claim victory.

Most serious conflict yet

The road toll conflict has been among the most serious to face Solberg PERSON since taking over as prime minister of a minority conservative government in 2013 DATE , and threatened to topple her government.

Higher road tolls and far more extensive toll systems set up in cities like Stavanger GPE ,
Bergen LOC and Oslo GPE left motorists faced with paying tens of thousands CARDINAL of more kroner per year, so it became a pocketbook issue that nurtured the rise of protest parties that have attracted large numbers of voters away from the established parties.

(a) BIO-format

We have sent two paragraphs from a longer text through the Stanford corenlp for Named Entity Recognition. The figure shows the result. (The break between sentence 3 and 4 is due to a typo in the source text we did not correct.)

A common format for representing NE-marked text is to use the so-called BIO-tags. Please present sentence 2 (“It was quickly … Friday evening.”) using the BIO-tag format and explain how the representation should be understood.

Fill in your answer here

Maximum marks: 10

(b) Evaluation

We have also run the same paragraph through Spacy. (For fairness sake, we should explain that we have used a small Spacy model; Spacy can do better with larger models.)

To compare the two, we will consider the Stanford corenlp result as the correct markup and the Spacy result as predicted outcome. Please calculate precision and recall for the various entity types. Explain how you find the numbers.

Also calculate the macro- and micro-averaged precision and recall across the entities.

For comparison, you may assume the following equivalences:

Corenlp	Spacy
PERSON	PERSON
ORGANIZATION	ORG
IDEOLOGY	NORP
DATE	DATE
TIME	TIME
MISC	MISC
NUMBER	CARDINAL
GPE	GPE
CITY	LOC

You can disregard the TITLE and P1Y entities, as we assume that Spacy does not aim at recognizing them.

You may ignore the additional detailed information in DATE, TIME, NUMBER.

If any expression ends up as 0/0, consider it undefined and exclude it from the (macro) averages.

Fill in your answer here

Maximum marks: 10

4(a) Softmax

A central formula in classification is based on the softmax function:

$$P\left(C_j \middle| \vec{x}\right) = \frac{e^{\overrightarrow{w_j}\cdot\vec{x}}}{\sum_{i=1}^{k} e^{\overrightarrow{w_i}\cdot\vec{x}}}$$

Explain the formula, in particular

1. What does $P(C_i \mid \vec{x})$ express in a classification task?
2. What is C_i?
3. What is $\vec{w}_i$?
4. What is $\vec{x}$?
5. What is k?

(Observe that $\vec{x}$ can also be written as bold face $\mathbf{x}$ and similarly with $\vec{w}_i$.)

Fill in your answer here

Maximum marks: 10

4(b) Recurrent Neural Nets

What is a recurrent neural net (RNN)? Sketch how a RNN can be used for POS-tagging. In particular, explain the softmax formula’s place in the model.

Fill in your answer here

Maximum marks: 10

5(a) Core dialogue concepts

Here is a short example of dialogue:

Person 1: *did you manage to finish the third obligatory assignment due yesterday?*

Person 2: *I sent it last week already!*

Person 1: *Impressive!*
Could you send me your solutions?

Person 2: *Sure, I'll send you the PDF*

Answer the following questions based on this dialogue:

1) What are the speech acts (according to Searle's taxonomy) associated with each of those 5 utterances? *(2 points)*

2) Do you observe some conversational implicatures? Briefly explain. *(2 points)*

3) Does the common ground of those two persons evolve in the course of this short interaction? Explain in 2-3 sentences. *(2 points)*

Fill in your answer here

Maximum marks: 6

5(b) Chatbot design

Assume you wish to develop a chatbot to answer questions about the current time in various locations around the world. The chatbot should for instance be able to respond to queries such as "*What is the current time in Buenos Aires?*", "*What is the time difference between Boston and Oslo?*" or "*If it is 6:00 AM in Oslo, what time is it in Tokyo?*".

Your first task is to decide what kind of chatbot development strategy you wish to follow. We have covered four alternative approaches during the course: handcrafted chatbots, IR-based chatbots, sequence-to-sequence chatbots and NLU-based chatbots.

Answer the questions below:

1) Which type of approach (among the four approaches above) do you think is most suitable for this task? Motivate your choice in one or two paragraphs. *(5 points)*

2) Which data will you need to collect to train/develop your chatbot, based on the approach chosen above? What type of annotations would you need to add to this data, if any? Explain in a few sentences. *(2 points)*

3) Which system modules (machine learning models or rule-based components) would you need to integrate in your chatbot? Describe in one or two paragraphs the general processing pipeline of your chatbot. *(4 points)*

4) How would you evaluate the performance of the resulting chatbot? Describe in one or two paragraphs the evaluation procedure you would follow. *(3 points)*

Fill in your answer here

Maximum marks: 14

5(c) Speech recognition

Answer the following questions:

1) Is is possible to have a Word Error Rate (WER) that is larger than 100%? Explain. *(2 points)*

2) During the lecture on speech processing, we mentioned that one challenge faced when training speech recognition models was the lack of explicit alignments between the speech inputs and the output transcriptions. What did we mean by that? Explain in 2-3 sentences. *(2 points)*

Fill in your answer here

Maximum marks: 4

5(d) Reinforcement learning

Assume you wish to develop a talking robot that can respond to various requests. When a human is perceived around the robot, the first first step is to engage the human, for instance with a greeting, accompanied or not by gestures. We wish to apply reinforcement learning to determine the best engagement strategy among two options: *SayHi* and *SayHiWithGestures*.

Formally speaking, we can frame this problem as an MDP with
- 2 states: *HumanNotEngaged* (which is the starting state) and *HumanEngaged* (which is a final state).
- 3 actions: *SayHi*, *SayHiWithGestures*, and *AskHowCanIHelpYou*. The first two actions (*SayHi* and *SayHiWithGestures*) are only available for the starting state *HumanNotEngaged*, while the action *AskHowCanIHelpYou* is only available for the final state *HumanEngaged*.

The transition model $P(s'|s,a)$ for this MDP is as follows:
- If the robot executes action *SayHi* in the state *HumanNotEngaged*, we have a probability 0.5 of reaching the state *HumanEngaged*, and a probability 0.5 of staying in the state *HumanNotEngaged.*
- If the robot executes action *SayHiWithGestures* in the state *HumanNotEngaged*, we have a probability 0.7 of reaching the state *HumanEngaged*, and a probability 0.3 of staying in the state *HumanNotEngaged.*

The reward function $R(s,a)$ of this MDP is as follows:
- The reward of executing action *SayHi* in the state *HumanNotEngaged* is -1
- The reward of executing action *SayHiWithGestures* in the state *HumanNotEngaged* is -2, since the physical gestures "cost" more to the agent in terms of energy and mechanical wear.
- Finally, the reward of executing action *AskHowCanIHelpYou* in the state *HumanEngaged* is +10. You can consider this action as the final one in this MDP, which means that the expected cumulative reward Q(*HumanEngaged, AskHowCanIHelpYou)* = 10.

Questions:
1) Compute the expected cumulative rewards Q(*HumanNotEngaged*, *SayHi*) and Q(*HumanNotEngaged*, *SayHiWithGestures*) based on the MDP described above. To compute those Q-values, you need to use Bellman's equation and refine your estimates through several iterations. You can stop after 5 iterations and use a discount factor γ=0.9. For the first iteration, you can initialize the Q-values to zero. *(10 points)*
2) Based on those Q-values, which engagement strategy should the robot chose? *(2 points)*

Tip: Bellman's equation is:
$Q(s,a) = R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a)\ max_{a'} Q(s',a')$

Fill in your answer here

Maximum marks: 12

6 Data privacy

In the field of data privacy, what is the difference between a direct identifier and a quasi-identifier? Explain in a few sentences, and illustrate those types of personal identifiers with a few examples.

Fill in your answer here

Maximum marks: 4