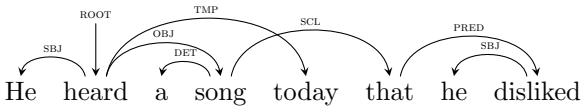
- **You may answer in English, Norwegian, Danish or Swedish.**
- **You should answer all questions. The weight of the various questions are indicated.**
- **You should read through the whole set to see whether anything is unclear so that you can ask your questions to the teachers when they arrive.**
- **If you think some assumptions are missing, make your own and explain them!**

1 Dependency syntax and parsing (25 %)

(a) Describe the formal conditions on well-formed dependency graphs.

- *connectedness*
- *acyclicity*
- *single-head*
- (*projectivity*)

(b) Consider the dependency graph for the English sentence *He heard a song today that he disliked*:



Is the graph above projective? Why or why not?
The graph is non-projective, since it has crossing branches.
Projectivity requires that if $i \rightarrow j$ then $i \rightarrow^* k$, for any k such that $i < k < j$ or $j < k < i$. This is not the case for the arc between *song* and *that*, since *today* has a head that is not a descendant of *song*.

(c) What are the main differences between graph-based and transition-based approaches to data-driven dependency parsing? What type of parser is Maltparser?

- graph-based: scoring of entire dependency graph, locating the highest scoring graph
- transition-based: predicting parse transitions, given parse history, constructing the optimal transition sequence
- Maltparser is transition-based

2 Semantic role labeling (25 %)

(a) Provide three roles from the traditional set of semantic roles along with a short description. Apply these to the sentence from exercise 1.

AGENT – participant that initiates the action, capable of acting with volition

PATIENT/THEME – participant that is affected by an action

EXPERIENCER – participant that is aware of the action described by the predicate, but who is not in control

Applied to sentence above:
[He]$_{EXP}$ heard a [song]$_{PATIENT}$ that [he]$_{AGENT}$ disliked

(b) Consider the following frame taken from FrameNet:

Definition:	
An Agent cuts an Item into Pieces using an Instrument:	
Frame Elements (core):	
Agent	person cutting the Item
Item	item being cut
Pieces	parts of the original Item
Frame Elements (non-core):	
Instrument	instrument with which the Item is cut
Manner	manner in which Item is cut
Place	where cutting takes place
Purpose	purpose for cutting
Lexical Units:	*carve, chop, cube, cut, dice fillet, mince, pare, slice*

(i) Briefly describe the notion of a *frame* and the different frame components above.
A frame is a background knowledge structure, defines a set of frame-specific semantic roles, so-called frame elements. Also includes a set of predicates that use these roles (the lexical units). Core elements are arguments, non-core are adjuncts.

(ii) In what way is Fillmore's frame semantics a response to criticism of the traditional set of semantic roles?
The traditional set of roles has been criticised for attempting to define a general set of roles that may be used for all predicates. As a response, Fillmore's notion of frame defines a set of roles only for a group of semantically related verbs, i.e. those relevant for a specific frame.

(c) In Semantic Role Labeling (SRL), parsing is often used as a pre-processing step for feature construction. Provide two examples of features commonly used in SRL that require syntactic parsing and use the dependency graph above to give one example for each of the features you describe.
Features that make use of syntactic analysis:

- Phrase type, e.g. NP, VP

- Path: from argument to predicate
- Head word
- Governing category: phrase-structure
- Dependency relation

For those that have dependency formulation, else N/A:

- Phrase type: N/A
- Path: e.g. from *song* to *heard*: OBJ ↑, or from *he* to *heard*: SBJ ↑ *pred* ↑ *scl* ↑ *obj* ↑
- Head word: *song*
- Governing category: N/A
- Dependency relation: SBJ and/or ROOT

3 Estimation (15%)

Kim has constructed a classifier and is testing it on a test set with 400 items. It achieves an accuracy of 0.64. Estimate a confidence interval for the accuracy of the classifier at level 0.99.

(A little help since you do not have a computer at hand: $\sqrt{0.64 \times 0.36} = 0.48$)

Formula for the interval is

$$[\hat{p} - Z^* \frac{s}{\sqrt{n}}, \hat{p} - Z^* \frac{s}{\sqrt{n}}]$$

Here

$n = 400$

$\hat{p} = 0.64$

$s = \sqrt{\hat{p}(1-\hat{p})} = \sqrt{0.64 \times (1 - 0.64)} = 0.48$

$Z^* = 2.576$

$$[\hat{p} - Z^* \frac{s}{\sqrt{n}}, \hat{p} - Z^* \frac{s}{\sqrt{n}}] = [0.64 - 2.576\frac{0.48}{\sqrt{400}}, 0.64 + 2.576\frac{0.48}{\sqrt{400}}]$$

$$= [0.64 - 2.576 \times 0.024, 0.64 + 2.576 \times 0.024]$$

$$\approx [0.64 - 2.5 \times 0.024, 0.64 + 2.5 \times 0.024] = [0.58, 0.70]$$

4 Collocations (20%)

Explain what a collocation is. There have been proposed several different measures for the association between the two words in a collocation. Describe two or three of them.

5 Machine learning (15%)

Multinomial logistic regression is a method for classification which is also called Maximum entropy classification. It is similar to Naive Bayes classification and can in some respects be considered a refinement of Naive Bayes classification. When trained on the same training material with the same set of features, the Logistic regression classifier should in principle perform at least as well as the Naive Bayes classifier when measured on the training material, and sometimes it will perform better. Without going deeply into the mathematical details, explain why this is so.

Naive Bayes (NB) and Multinomial logistic regression (MLR) are so-called log-linear classifiers. If an NB and an MLR classifier over the same data use the same features, the only difference between the two is how the features are weighed.

The NB classifier weigh each feature according to how often it co-occurs with the different classes in the training set. There is only one way to select the weights.

The MLB classifier selects weights according to how well a classifier using these weights classify the training data. In principle it may assign any set of weights including the weights assigned by the NB classifier. Hence it will perform at least as well as the NB classifier.

In case there are other weights that perform better than the weights assigned by the NB classifier, an MLR classifier may select them and perform better than the NB classifier.

Also, explain why adding more features may deteriorate the results of the Naive Bayes classifier, but not of the Logistic regression classifier.

Suppose an NB classifier uses optimal weights and we add a new feature f_i. Assume further that this feature has exactly the same distribution with respect to classes as a feature already used, say f_j. The effect of adding f_i and assigning it a weight w_i on the basis of its distribution will not alter any of the other weights. The net effect will be the same as if we did not add any new features but instead doubled the weight of f_j and kept the other weights untouched. If the original weights were optimal, they may be suboptimal after such a change.

The MLR on the other hand may assign the weight $w_i = 0$ and keeping the results unaltered. (It may of course also alter the weights on the other features if that had been beneficial).

In spite of this, Jurafsky and Martin claim "Furthermore, naive Bayes works extremely well (even better than logistic regression or SVMs) on small datasets or short documents." Why do you think that is the case?

Since the MLR classifier adjust its weights very carefully to fit the training data, there is a chance of overfitting, in particular if the dataset is small or the feature set is large.

Since the NB classifier assign weights on the basis of co-occurence of features and classes, that property is more robust when transferring from training data to other data, hence the overfitting may be less.

END