Statistical Natural Language Processing
Assignment 8

Sven Stauden 2549696
Janis Landwehr 2547715
Carsten Klaus 2554140

Full points!

Exercise 1

The performance of the base line system highly depends on the number of likely senses. Therefore, words that have several sensen, where in general only one sense is used, the base classifier will perform very well. This drastically changes, when a word has e.g. two or three often occuring meanings. For example, the word dog has the meanings 'Hund' and 'Bauklammer'. However, the second meaning almost never occurs and thus the base classifier will perform well by assigning the class 'Hund' to all meanings. An example, where the base classifier performs poorly would e.g. be the example *bank* from the lecture, which has several often occuring meanings.

Exercise 2

a) Bayes Classifier only considering one feature

For this approach one could design the classification problem as in the following:

$$s_k = \underset{w_c \in C, s_k}{argmax} P(s_k|w_c)$$

Here we consider all possible indicator words w_c in the contest C individually and compute the most probable word meaning s_k. Finally we choose the indicator word which delivers the highest certainty over a word meaning.

b) Information-theoretic method with many features

A variation for this approach would be to replace the partitioning Q_1, Q_2 of possible indicator words with considering a context window around the ambiguous word. One maximizes this mutual information by adapting different partitioning combinations for P. The best partition indicates the word meaning with the highest correlation.

$$\underset{P}{argmax}\ I(P;C) = \underset{P}{argmax} \sum_{xf \in C} \sum_{t \in P} p(xf,t) log \left(\frac{p(xf,t)}{p(xf)p(t)} \right)$$

Exercise 3

Unsupervised learning techniques also suffer from overfitting (adjusting the model to the data noise). Therefore, by e.g. choosing a high number of clusters, the model will eventually overfit. This can also happen by having a unbalanced data set. Evaluating on the training set will lead to a high accuracy, while the model will nevertheless perform poorly on unseen data.

Exercise 4

In Lesk's original algorithm the similarity value between the considered word description $D(s_x)$ and the merging of the lexicon definitions of all words in the considered context $\cup E(v_j)$, with $v_j \in C$ gets computed.

In this simplified version we skip considering the lexicon definitions of all context words but only consider the words $v_j \in C$ itself. This reduces the amount of used data drastically but also reduces the chance of finding important word correlations in the context.

Exercise 5

see source code

Exercise 6

see source code

Exercise 7 and 8

```
bass
Accuracy (default) 0.5887850467289719
Accuracy most frequent sense 0.9065420560747663
```

```
Accuracy (jaccards) 0.5887850467289719
crane
Accuracy (default) 0.8315789473684211
Accuracy most frequent sense 0.7578947368421053
Accuracy (jaccards) 0.8315789473684211
motion
Accuracy (default) 0.6716417910447762
Accuracy most frequent sense 0.7064676616915423
Accuracy (jaccards) 0.6716417910447762
palm
Accuracy (default) 0.8009950248756219
Accuracy most frequent sense 0.7114427860696517
Accuracy (jaccards) 0.8009950248756219
plant
Accuracy (default) 0.5957446808510638
Accuracy most frequent sense 0.5425531914893617
Accuracy (jaccards) 0.5957446808510638
tank
Accuracy (default) 0.746268656716418
Accuracy most frequent sense 0.373134328358209
Accuracy (jaccards) 0.746268656716418
```

The accuracy between Lesks algorithm and the most frequent sense highly depends on the given word. Words that have only one very common word seem to work relatively well on the most frequent sense. However, words that have several likely meenings (e.g. tank) perform quite bad, while Lesks algorithm has a good accuracy.
Comparing the two similarity measures, one can see that both perform equally well on the classification tasks. This is due to their monotonic relationship. IN the next task we will show this.

Proof

1)

$$\begin{aligned}
\frac{S}{2-S} &= \frac{\frac{2|X\cap Y|}{|X|+|Y|}}{2-\frac{2|X\cap Y|}{|X|+|Y|}} \\
&= \frac{\frac{2|X\cap Y|}{|X|+|Y|}}{\frac{2|X|+|Y|}{|X|+|Y|}-\frac{2|X\cap Y|}{|X|+|Y|}} \\
&= \frac{2|X\cap Y|}{|X|+|Y|}\cdot\frac{|X|+|Y|}{2(|X|+|Y|-|X\cap Y|)} \\
&= \frac{|X\cap Y|}{|X|+|Y|-|X\cap Y|} \\
&\overset{*}{=} \frac{|X\cap Y|}{|X\cup Y|} = J
\end{aligned}$$

2)

$$\begin{aligned}
\frac{2J}{1+J} &= \frac{\frac{2|X\cap Y|}{|X\cup Y}}{1+\frac{|X\cap Y|}{|X\cup Y|}} \\
&= \frac{\frac{2|X\cap Y|}{|X\cup Y}}{\frac{|X\cup Y|}{|X\cup Y|}+\frac{|X\cap Y|}{|X\cup Y|}} \\
|| &= \frac{2|X\cap Y|}{|X\cup Y}\cdot\frac{|X\cup Y|}{|X\cup Y|+|X\cap Y|} \\
&= \frac{2|X\cap Y|}{|X\cup Y|+|X\cap Y|} \\
&\overset{*}{=} \frac{2|X\cap Y|}{|X|+|Y|} = S
\end{aligned}$$

Finally we want to show (*) namely that $|A\cup B| = |A|+|B|-|A\cap B|$
For full proof, see `http://pages.cs.wisc.edu/~dieter/Courses/2011s-CS240/Homework/solutions02.pdf`, Problem 1.
Idea:
$A = A\setminus B$
$B = B\setminus A$
Disjoint Union results in $A\cup B = (A\setminus B)\cup(B\setminus A)\cup(A\cap B)$
We can use set laws and algebra to proof that $|A\cup B| = |A|+|B|-|A\cap B|$