Lecture Speechprocessing

Exercise 6

1. Language Modelling
   (a) What is a language model, which differ bi- and trigrams?
   (b) What is a perplexity?
   (c) What must be taken into account during the creation of language models?
   (d) For which systems language models can senseless?

2. Consider the following training corpus:
   ich schaue zum himmel. die sonne brennt und blendet. die luft ist heiß weil die sonne seit tagen vom himmel brennt. die haut brennt vom sand und brennt vom öl am strand.
   (a) How big is the vocabulary V?
   (b) How Many bi- and trigrams are theoretically possible?
   (c) Calculate the Zero-gram probabilities
   (d) Calculate now using Zero, uni- and bi-grams, the probability of occurrence of the following test corpus:

   die sonne brennt vom himmel

   (e) What is the perplexity of the test corpus for Zero, uni- and bi-grams?

3. A Corpus contains the vocabulary **A**, **B** and **C**. During trianing the word **A** occurs 4 times, **B** 3 times and **C** 0 times.
   (a) What is the probability of the sequence of letters $ABABBCAAABAAB$ when using unigrams?
   (b) What is the perplexity?
   (c) How do the unigram probabilities change when a Jeffrey smooting is used?

4. Language models: Good-Turing Estimation The *Austen* Text Corpus has a vocabulary of $V = 14,585$ words. A total of $N = 617,091$ words are contained inthe Corpus, while $eta = 199,252$ bigrams are seen. In the table means η_r: bigrams, that have been seen r-time in the Corpus.
   (a) How Many unseen bigrams are there?
   (b) Calculate and interpret r^* for $r = 0, 2, 7, 9, 843$.
   (c) What is the bigram probability is for *she was* and *both sisters* without smoothing?
   (d) Calculate using the following table, the bigram probabilities for the sentence *she was inferior to both sisters* according to the Good-Turing Method (The Predecessor of *she* is *person*).

r	η_r	r	η_r
1	138741	8	1342
2	25413	9	1106
3	10531	10	896
4	5997	...	
5	3565	843	1
6	2486	844	0
7	1754	...	

w	$\#(w)$	w_1w_2	$\#(w_1w_2)$
person	223	person she	2
she	6917	she was	843
was	9409	was inferior	0
inferior	33	inferior to	7
to	20042	to both	9
both	317	both sisters	2