Lehrstuhl für Informatik 6 **June 2, 2015**
Rheinisch–Westfälische Technische Hochschule Aachen **Andreas Guta**
Prof. Dr.–Ing. H. Ney

6. Exercise Sheet **Statistical Methods in Natural Language Processing**

The solutions to the problems may be submitted until **Tuesday, June 9, 2015** before the exercise lesson. Upload a digital version to the L2P and bring your written/printed solution to the exercise lesson. Condition for obtaining the **Leistungsnachweis** (*Schein*) "Statistical Methods in Natural Language Processing" is the successful solution of 50% of the problems and the presentation of the solution of at least two problems in the exercise lessons.
The solutions to the problems can be submitted in groups of up to **two** students.

1. [2 Points] Consider you have a set of (bigram) language models $\{p_k(w|v)\}_{k=1}^K$ and consider the two ways of combining them

$$p(w_1^N) = \sum_{k=1}^{K} \left[\lambda(k|w_1^N) \prod_{n=1}^{N} p_k(w_n|w_{n-1}) \right]$$

and

$$p(w_1^N) = \prod_{n=1}^{N} \sum_{k=1}^{K} \lambda(k|w_1^n) p_k(w_n|w_{n-1}) \,,$$

with $\lambda()$ a probability distribution. Explain the main difference between the two models and discuss advantages and disadvantages, taking applications into account.

2. [2 Points] Consider applying the absolute discounting smoothing method to the multinomial distribution, as used in the text classification task. Write the leaving-one-out log-likelihood function considering the variations "leaving-one-word-out" and "leaving-one-document-out". What would be good choices for the generalized distribution β? (intuitively, no need to take the derivative)

3. [1 Point] What is the problem with the following concept for a bigram-based POS tagger?

   "For each position n, consider the neighbouring positions $n-1$ and $n+1$ and find the best decision for position n"

   A simple (textual) explanation is sufficient.

4. [3 Points] Many applications of dynamic programming (DP) in speech and natural language processing can be formulated as follows:

   We are given a local cost function $h_n(u_{n-1}, u_n) \geq 0$ depending on a position (or time) $n = 1, \ldots, N$ and decisions u_n taken at position n and decision u_{n-1} at position $n-1$. The decisions are from a finite set $u \in \{1, 2, \ldots, U\}$. The optimization problem is

$$\arg\min_{u_1^N} \left\{ \sum_{n=1}^{N} h_n(u_{n-1}, u_n) \right\} .$$

   a) Give the DP recursion equations for this problem. What is the time and memory complexity? Which type of dependence in the cost function is that: zero-, first-, second-order?

b) Assume the following type of cost function for any $n = 1, \dots, N$:

$$h_n(u', u) = \begin{cases} h_n(u) & \text{if } |u - u'| \leq 1 \\ \infty & \text{otherwise} \end{cases} .$$

Give the DP recursion and its time and memory complexity. Which type of dependence in the cost function is that: zero-, first-, second-order?

c) What would be a zero-order dependence in the cost function? What simplifications could be done in this case?