Aalto University School of Science, Department of Information and Computer Science;
Krista Lagus, Timo Honkela, Oskar Kohonen, Mikko Kurimo, Kalle Palomäki, Jaakko Peltonen, Juho Rousu, Teemu Ruokolainen, Sami Virpioja

.

T-61.5020 Statistical processing of natural language, exam 26 April 2014

Please, indicate the following information in each paper of your answer:

- name, student number, faculty)
- phrase: "T-61.5020, exam 26.4.2014"

In the evaluation of essay answers, attention is paid to the punctuality and clarity of your answer. Avoid overly long answers. You can also answer in *Finnish.*

1. Explain shortly (2-3 sentences) each of the following terms (1p/term).
   a) N-gram language model
   b) Chomsky hierarchy
   c) Context dependent HMM
   d) Crowdsourcing
   e) Vector space model
   f) Viterbi algorithm
   g) Random projection
   h) Named entity recognition
   i) Kernel trick
   j) Subsequence kernel

2. According to the Bayes rule, the probability of a particular meaning s_k is obtained, when the context c is known, using the equation

$$P(s_k|c) = \frac{P(c|s_k)P(s_k)}{P(c)}$$

   Naive Bayes method is based on the assumption that the occurrence probabilities of words in the context do not depend on each other.

   (a) How could you apply this equation in the disambiguation task? (Explain what is disambiguation.) (3p)

   (b) Outline shortly an unsupervised approach for modeling ambiguity of words. (3p)

3. Compare similarities and differences between a Part-Of-Speech (POS) tagger, a morphological analyzer (such as the OMorFi toolkit for Finnish) and a system for morpheme segmentation based on unsupervised learning (such as Morfessor).

   a) Describe the basic objectives of each method/tool (2p).

   b) What is the role of explicit tagging in each approach? (1p)

   c) What kind of statistical machine learning is or can be involved in each of the tasks? (3p)

4. Google has developed a system that translates spoken sentences expressed in one language into their counterparts in another language. For instance, if one says in German "Ich möchte einige Kaiserschmarren haben", the system replies in English "I want to have

some pancake". A Finnish sentence with similar meaning "Haluaisin pannukakkua" is translated into a French expression "Je voudrais crêpe". The system takes spoken sentences as input and provides spoken sentences as output. In addition, it shows the written counterparts to the user.

Google is known to strongly rely on statistical machine learning methods and data-driven approaches in their system development. The number of languages included in the system is close to one hundred.

Present an outline of the overall architecture of Google's system described above (2p). In particular, show what kind of basic modules are involved in speech recognition and machine translation when conducted using statistical machine learning (3p). Describe in additional detail one particular task or method (such as translation model, language model, acoustic model, HMM, etc.) and discuss what kind of challenges the developers are likely to encounter in the development, for instance, in the manner how available data is utilized (3p).