Aalto University School of Science and Technology, Faculty of Information and Computer Science; Timo Honkela (tel. 050 384 1578), Mikko Kurimo, Jaakko Väyrynen, Sami Virpioja and Oskar Kohonen

T-61.5020 Statistical processing of natural language, exam 20 May 2011

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

- name, student number, faculty)
- phrase: "T-61.5020, exam 20.5.2011"

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).
   - equivalence class
   - validation set
   - collocation
   - decoding in speech recognition
   - stemming
   - recall and precision (in IR)
   - audio information retrieval
   - morpheme discovery

2. Explain shortly (5-10 sentences) each of the following terms (2p/term).
   - part-of-speech (POS) tagging
   - smoothing/discounting methods

3. You are looking for an answer to a problem that horsemen have pondered for a long time: "Varför får hästen inte gå i bastun?" ("Why is a horse not allowed in sauna"). The solution is known only to the Swedish: "Den blir ren och äter laven". You have a language model and translation probabilities between English and Swedish words in Table 1. You have two strong candidates for the translated sentence:

   - "It becomes clean and eats the seats"
   - "It turns into a reindeer and eats lichen"

   Which translation is more probable and why? (6p).

   You can use:
   $\hat{e} = \underset{e}{\text{argmax}}\, P(e|s) = \underset{e}{\text{argmax}}\, P(e)P(s|e)$

   and

   $P(s|e) = \frac{1}{Z} \sum_{a_1=0}^{l} \dots \sum_{a_m=0}^{l} \prod_{j=1}^{m} P(s_j|e_{a_j})$

   where m is the length of the original Swedish sentence, l is the length of the translated English sentence, and a_j are the possible word (or phrase) alignments.

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

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

w	$P(w)$
it	0.18
becomes	0.05
clean	0.01
eats	0.1
the	0.12
seats	0.02
turns	0.07
into	0.11
a	0.21
reindeer	0.01
and	0.13
lichen	0.01

w_1	w_2	$P(w_1 \rightarrow w_2)$
it	den	1.0
becomes	blir	0.7
becomes	klär	0.3
turns	blir	0.7
turns	vänder	0.3
into	[]	1.0
clean	ren	0.9
clean	städa	0.1
a	[]	1.0
reindeer	ren	1.0
and	och	1.0
eats	äter	1.0
the	[]	1.0
seats	laven	0.1
seats	stolar	0.9
lichen	laven	1.0

Table 1: Unigram model on the left, translation probabilities on the right.

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

a) Using this assumption, determine the equation of a naive Bayesian classifier. (2p)

b) How could you apply this equation in the disambiguation task? (Explain what is disambiguation.) (4p)

5. A mobile phone company wishes to develop a method for dictating a text message in a spoken form using their own mother tongue (e.g. Japanese or Russian). The speech input is transformed into text and then the message is automatically translated into recipient's language (e.g. German or Hindi). The recipient receives the message in a written form.

Present the overall architecture of the system with components based on *statistical methods* (2p). In particular, present a careful but concise description of one *central* method used in speech recognition (2pm) and another used in statistical machine translation (2p). Explain shortly why these methods are useful in these tasks (2p).