T-61.5020 Statistical Natural Language Processing

Exercises 9 — Statistical machine translation
Version 1.1

1. You are looking for answer to a problem that horsemen have pondered for a long time: “Varför får hästen inte gå i bastun?” (“Why shouldn’t horse go in sauna?”). The solution is known only by the Swedish: “Den blir ren och äter laven”. You have a language model and translation probabilities between English and Swedish words given 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 is the more probable one?

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 in left, translation probabilities in right.*

2. (Computer assignment) Let’s examine the problems of estimating translation probabilities. European Parliament Proceedings Parallel Corpus[1] consists of sentence aligned texts between pairs of various European languages. Use a suitable parallel file, e.g. Finnish-English[2].

   The corpora include XML-style tags and other information not needed here. They can be removed using a Python script available in the course’s web page[3]. The corpus

[1] *Europarl: A Parallel Corpus for Statistical Machine Translation*, Philipp Koehn, MT Summit 2005. `http://www.statmt.org/europarl/`

[2] `http://www.statmt.org/europarl/v2/fi-en.tgz`

[3] Address: `http://www.cis.hut.fi/Opinnot/T-61.5020/Exercises08/extra/cleanfile.py`
Usage example: `python cleanfile.py corpus_in corpus_out`

package has separate files for the two languages, and the same line numbers of the same files are the corresponding sentences.

Next choose a relatively common word f of source language, e.g. Finnish. Find all the sentences which include that word from the Finnish corpus. Then go through the target language (e.g. English) and collect all the words e that are in the corresponding sentences (lines) where the Finnish word was found, together with their co-occurrence counts ($C(e, f)$). Then try to find the most probable translation(s) from this set of words.

a) Start using directly the number of co-occurrences $C(e, f)$. How does it work?

b) Try then to weight the values by the number of sentences where e occurs in the whole corpus, $C(e)$.

c) Try other kind of weights and/or statistical methods to find the correct translation possibilities. Some ideas can be found from the Exercise 5 that concerned collocations.