T-61.5020 Luonnollisten kielten tilastollinen käsittely

Harjoitus 9, ke 28.3.2007, 12:15–14:00 — Tilastollinen konekääntäminen
Versio 1.0

1. Etsit ratkaisua hevosmiehiä pitkään pohdituttaneeseen ongelmaan, "Varför får hästen inte gå i bastun?". Ratkaisun ongelmaan tuntevat vain ruotsalaiset ("Den blir ren och äter laven"). Osaat englantia ja käytössäsi on sekä taulukon 1 kielimalli ja käännöstiedot. Sinulla on kaksi vahvaa ehdokasta vastauksen käännökseksi:

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

   Kumpi on todennäköisempi?

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

Taulukko 1: *Vasemmalla unigrammikielimalli, oikealla käännöstodennäköisyydet.*

2. (Tietokoneharjoitus) Tutustutaan hieman käännöstodennäköisyyksien estimoinnin ongelmiin. Käytetään aineistona europarlamentin istuntojen pöytäkirjoista tehtyä lausetasolla kohdistettua rinnakkaiskorpusta[1].

   Valitaan rinnakkaiskorpuksista suomi-englanti -aineisto[2]. Tekstit sisältävät XML-tyylisiä tageja ynnä muuta "turhaa" informaatiota, jotka siivotaan pois. Kurssin sivuilta löytyy valmis Python-ohjelma tätä varten[3]. Aineistossa on erilliset tiedostot englannin- ja suomenkielisille lauselle, ja samassa tiedostossa samalla rivinumerolla olevat lauseet vastaavat toisiaan.

[1] *Europarl: A Multilingual Corpus for Evaluation of Machine Translation*, Philipp Koehn, Draft, Unpublished. `http://www.statmt.org/europarl/`

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

[3] Osoitteessa `http://www.cis.hut.fi/Opinnot/T-61.5020/Laskarit07/scripts/cleanfile.py`
Käyttö esim.: `python cleanfile.py tiedosto_sisään tiedosto_ulos`

Valitse seuraavaksi suhteellisen yleinen aineistosta löytyvä suomenkielinen sana (f), esimerkiksi "tosiasia". Etsi suomekielisistä teksteistä kaikki lauseet joissa sana esiintyy, ja kerää vastaavista englanninkielisistä lauseista mahdolliset käännössanat (e), sekä jokaiselle sanalle esiintymien yhteismäärä niissä lauseissa, joissa valittu suomenkielinen sana esiintyi ($C(e, f)$).

a) Miltä käännös vaikuttaa, jos valitset sanan suoraan luvun $C(e, f)$ perusteella?

b) Entä jos painotat lukua e:n kokonaisesiintymien määrällä $C(e)$?

c) Kokeile muita erilaisia painotuksia tai tilastollisia testejä hyvien käännösten löytämiseksi. Vinkkejä saa esimerkiksi laskuharjoituksesta 5 (kollokaatiot).