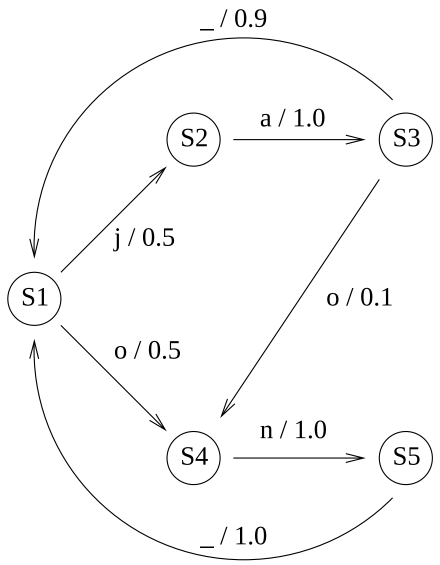
T-61.5020 Luonnollisten kielten tilastollinen käsittely

Harjoitus 11, ke 18.4.2007, 12:15–14:00 — Puheentunnistus ja kielimallien evaluointi
Versio 1.0

1. Tarkastellaan yksinkertaista äännemallia, joka on esitetty kuvassa 1. Mallissa on viisi tilaa, $S_1, \ldots, S_5$, joista S_1 on sekä alku- että lopputila. Jokaiselle tilojen väliselle kaarelle on annettu siirtymätodennäköisyys $a_{ij} = P(S_j|S_i)$ sekä kaarta vastaava äänne. Alkutilaan johtavat kaaret vastaavat sananväliä, ja ovat niin sanottuja nollatransitioita, jotka eivät tuota havaintoja. Muille kaarille on estimoitu jakauma havaintotodennäköisyyksille $b_{ij}(o_k) = P(o_k|S_i \to S_j)$.



Kuva 1: Äännemalli.

Puhesignaalista on saatu laskettua piirrevektorit $o_1, \ldots, o_4$. Taulukossa 1 on kaarien jakaumasta lasketut havaintotodennäköisyydet kunkin vektorin kohdalle.

Taulukko 1: Havaintotodennäköisyydet $b_{ij}(o_k)$.

i, j	o_1	o_2	o_3	o_4
1,2	10^{-1}	10^{-2}	10^{-3}	10^{-3}
2,3	10^{-3}	10^{-1}	10^{-1}	10^{-3}
3,4	10^{-3}	10^{-1}	10^{-1}	10^{-4}
4,5	10^{-3}	10^{-4}	10^{-3}	10^{-1}
1,4	10^{-3}	10^{-2}	10^{-2}	10^{-4}

a) Laske todennäköisin tilajono havainnoille viterbi-algoritmin avulla. Tilajonon pitää loppua ja alkaa tilasta S_1. Mikä sana tai sanajono tilasiirtymistä muodostuu?

b) Käytetään tunnistuksessa apuna kielimallia. Tehtävän kannalta relevantit kielimallitodennäköisyydet ovat seuraavat:

$$\begin{array}{lllllllll} P(\mathrm{ja}) & = & 10^{-2} & P(\mathrm{ja}\,|\,\mathrm{ja}) & = & 10^{-4} & P(\mathrm{ja}\,|\,\mathrm{on}) & = & 10^{-2} \\ P(\mathrm{on}) & = & 10^{-2} & P(\mathrm{on}\,|\,\mathrm{ja}) & = & 10^{-2} & P(\mathrm{on}\,|\,\mathrm{on}) & = & 10^{-4} \\ P(\mathrm{jaon}) & = & 10^{-5} & & & & & & \end{array}$$

Laske todennäköisin tilajono ja sitä vastaava(t) sana(t). Parhaat polut täytyy laskea jokaiselle sanahistorialle erikseen. Kerro kielimallitodennäköisyys mukaan heti sanan valinnan yhteydessä.

2. Erilaisten ja varsinkin erilaisia yksiköitä käyttävien kielimallien vertailu ei ole täysin suoraviivaista. Siispä tutkitaan tarkemmin miten se onnistuu.

   Opetusaineistosta on opetettu kaksi erilaista tilastollista sanojan pilkontaa morfeiksi, A ja B. Samasta aineistosta on opetettu kolme eri kokoista kielimallia kummallekin pilkonnalle. Koot ovat mallien sisältämien n-grammien määriä. Erillisestä sadantuhannen sanan testiaineistosta on laskettu kaikille malleille risti-entropiat *yksikköä kohti*. Tulokset on esitetty taulukossa 2.

Taulukko 2: Entropiatulokset. Testidatassa oli 100 000 sanaa.

	Yksiköitä		Risti-entropia 1		Risti-entropia 2		Risti-entropia 3	
Pilkonta	tyyppejä	testidatassa	H_M	koko	H_M	koko	H_M	koko
A	2 114	344 960	4.54	472 227	4.39	664 601	4.31	998 907
B	6 535	301 271	5.19	518 286	5.02	712 133	4.93	1 049 750

Lisäksi kielimalleja testataan puheentunnistusjärjestelmässä. Tunnistustuloksista lasketaan virheellisesti tunnistettujen sanojen osuus (word-error-rate, WER). Luvut ovat taulukossa 3.

Taulukko 3: Puheentunnistustulokset.

	Tunnistus 1		Tunnistus 2		Tunnistus 3	
Pilkonta	mallin koko	WER	mallin koko	WER	mallin koko	WER
A	472227	17.64	664601	15.04	998907	14.25
B	518286	17.54	712133	15.01	1049750	13.97

Selvitä annettujen tulosten perusteella kumpi malleista vaikuttaa toimivan paremmin entropiatestien mukaan? Entä tunnistuskokeiden valossa? Kuinka luotettavina johtopäätöksiä voi pitää?