T-61.281 Luonnollisen kielen tilastollinen käsittely

Harjoitus 2, ti 3.2.2004, 8:30-10:00 – Entropia, hämmentyneisyys, kontekstivapaa kieli, Versio 1.0

1. Otetaan pieni kieli johon kuuluu kahdeksan sanaa:

W	P(W)
'kissa'	$\frac{3}{32}$
'tuuli'	$\frac{3}{16}$
'kiipeilijä'	$\frac{7}{32}$
'naukaisi'	$\frac{1}{8}$
'tuivertaa'	$\frac{1}{8}$
'katosi'	$\frac{1}{4}$

   a) Oletetaan että lähde tuottaa satunnaismuuttujan X arvoja yo. taulukon todennäköisyyksien mukaan. Mikä on lähteen entropia $H(X)$?

   b) Tarkemmin asiaa tutkittaessa käykin ilmi että kielessä on lauserakenne 'SV' jossa kategoriat S $\in$ {'kissa', 'tuuli', 'kiipeilijä'} ja V $\in$ {'naukaisi', 'tuivertaa', 'katosi'}. Satunaismuuttujien yhteistodennäköisyysjakauma P(S,V) on:

	'naukaisi'	'tuivertaa'	'katosi'	
'kissa'	$\frac{1}{8}$	0	$\frac{1}{16}$	$\frac{3}{16}$
'tuuli'	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{1}{16}$	$\frac{3}{8}$
'kiipeilijä'	$\frac{1}{16}$	0	$\frac{3}{8}$	$\frac{7}{16}$
	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	

   Mikä on lähteen entropia, kun tiedetään, että edellinen symboli kuului joukkoon S, eli mikä on $H(X_i|X_{i-1} \in S)$?

2. Tarkastellaan edellisesta laskarista tuttua satunnaista kieltä: 30 symbolia, jotka kukin ovat yhtä todennäköisiä. Yksi symboleista on sanaväli.

   a) Lähde generoi merkkejä yhden kerrallaan. Mikä on lähteen entropia ?

   b) Lähde generoi yhden sanan kerrallaan (eli merkki/merkkejä + sanaväli). Kutakin sanaa kohdellaan yhtenä omana kokonaisuutenaan. Mikä on tällaisen lähteen entropia ?

3. Meillä on kolme kielioppia, jotka on esitetty seuraavan sivun taulukoissa. Testimateriaalina meillä on kaksi lausetta:

   a) Kissa menee puuhun.

   b) Valas on kala paitsi ettei.

   Laske kunkin mallin hämmentyneisyys (perplexity) molemmille testilauseille. Ovatko tulokset keskenään vertailukelpoisia ?

   Hämmentyneisyys voidaan määritellä testijoukon sanojen todennäköisyyksien geometrisen keskiarvon käänteislukuna:

$$Perp(w_1, w_2, \ldots, w_n) = P(w_1, w_2, \ldots, w_n)^{-\frac{1}{n}}$$

Malli 1	Malli 2
P(sana='kissa')=0.1	P(sana=subjekti)=0.33
P(sana='koira')=0.1	P(sana=verbi)=0.33
P(sana='valas')=0.1	P(sana=kohde)=0.33
P(sana='kala')=0.1	
P(sana='istui')=0.1	
P(sana='menee')=0.1	
P(sana='on')=0.1	
P(sana='puuhun')=0.1	
P(sana='kuuhun')=0.1	
P(sana='suuhun')=0.1	

Malli 3	
P(sana='kissa' | sana=ensimmäinen)	=0.25
P(sana='koira' | sana=ensimmäinen)	=0.25
P(sana='valas' | sana=ensimmäinen)	=0.25
P(sana='kala' | sana=ensimmäinen)	=0.25
P(sana='istui' | edellinen_sana ∈ {'kissa','koira','valas','kala'})	=0.33
P(sana='menee' | edellinen_sana ∈ {'kissa','koira','valas','kala'})	=0.33
P(sana='on' | edellinen_sana ∈ {'kissa','koira','valas','kala'})	=0.33
P(sana='puuhun' | edellinen_sana ∈ {'istui','menee','on'})	=0.33
P(sana='kuuhun' | edellinen_sana ∈ {'istui','menee','on'})	=0.33
P(sana='suuhun' | edellinen_sana ∈ {'istui','menee','on'})	=0.33

4. Meillä on seuraavanlainen kielioppi (e tarkoittaa tyhjää merkkiä):

$$
\begin{array}{lcl}
L & \rightarrow & SubOb\ V \\
V & \rightarrow & Verb\ SubOb \\
SubOb & \rightarrow & \begin{cases} Kuvaus\ Subst \\ Kuvaus \end{cases} \\
Kuvaus & \rightarrow & \begin{cases} e \\ Adj \\ Kuvaus\ Gen\ Kuvaus \end{cases} \\
Adj & \rightarrow & \begin{cases} \text{'keltaisen'} \\ \text{'vikkelä'} \\ \text{'rapistunut'} \end{cases} \\
Subst & \rightarrow & \begin{cases} \text{'jänis'} \\ \text{'karmi'} \\ \text{'laatikkoon'} \end{cases} \\
Verb & \rightarrow & \begin{cases} \text{'oli'} \\ \text{'hyppäsi'} \end{cases} \\
Gen & \rightarrow & \text{'oven'}
\end{array}
$$

Rakenna jäsennyspuu seuraaville lauseille käyttäen yllä annettua kielioppia.

a) Keltaisen oven karmi oli rapistunut.

b) Vikkelä jänis hyppäsi laatikkoon.

Keksi yksi järjetön lause, joka noudattaa ylläolevaa kielioppia. Sanalistoihin voi lisätä sanoja. Keksi yksi järkevä lause, jota kielioppi ei osaa jäsentää, vaikka se tuntisikin riittävän määrän sanoja.