Introduction to Symbolic and Statistical NLP in Scheme

Damir Ćavar
dcavar@unizd.hr

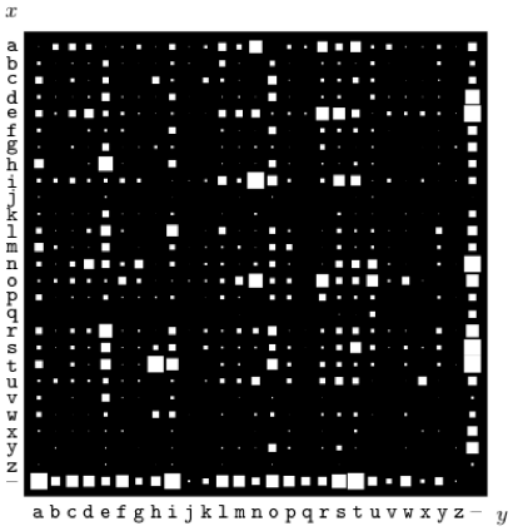
ESSLLI 2006, Malaga

July/August 2006

Frequency Profiles

- *Uni-* and *Bi-gram* frequencies . . .
- General *n*-gram models
- Examples: from [MacKay(2003)]

i	a_i	p_i
1	a	0.0575
2	b	0.0128
3	c	0.0263
4	d	0.0285
5	e	0.0913
6	f	0.0173
7	g	0.0133
8	h	0.0313
9	i	0.0599
10	j	0.0006
11	k	0.0084
12	l	0.0335
13	m	0.0235
14	n	0.0596
15	o	0.0689
16	p	0.0192
17	q	0.0008
18	r	0.0508
19	s	0.0567
20	t	0.0706
21	u	0.0334
22	v	0.0069
23	w	0.0119
24	x	0.0073
25	y	0.0164
26	z	0.0007
27	–	0.1928



Frequency Profiles

- What can we do with *n*-gram frequency profiles?
  - Compression, modeling expectations, study of quantitative language properties (e. g. also dialectal or cross-linguistic variation), . . .
- Develop applications
  - What value for *n* is best for what purpose?

Frequency Profiles

- Over large corpora lexical frequencies divide lexical classes (more or less well).
- Over individual texts lexical frequencies identify topics or document classes (more or less well).
- In general, n-gram models over orthographic or phonemic/phonetic representations are language specific (express morphological or phonotactic regularities).

Collocations

- Hypothesis testing (significance tests), e. g. χ^2

  $\chi^2 = \sum \frac{(observed - expected)^2}{expected}$

- Mutual Information

  $I(x;y) = P(x,y) lg \frac{P(x,y)}{P(x)P(y)}$

Language Identification

- *N*-gram models for Language Identification
- Files: `lid.ss`, `trigraph.ss`, `lang-*.ss`
- Calculations:
  - Mean of frequencies
  - Deviation

Language Identification

- Generate a persistent trigram model for each language
- Generate a trigram model for some unknown text
- Task:
  - For each language model, for each trigram from the unknown text, calculate the sum of the absolute difference between the trigrams in the two models.
  - The smallest absolute difference will be found between texts from the same language.

References

[MacKay(2003)] David J. C. MacKay. *Information theory, inference, and learning algorithms*. Cambridge University Press, Cambridge, UK; New York, 2003.