Using association statistics to identify fixed expressions

Begoña Villada Moirón
Alfa-Informatica
villada@let.rug.nl

February 21, 2005

Overview

- Problem statement
- What do I want to find out?
- Methodology
- Population object of study
  - ⋆ categorical variables
  - ⋆ data sampling
  - ⋆ contingency table
- Measuring associations between variables
  - ⋆ pointwise mutual information

- ⋆ Pearson's χ^2
- ⋆ log-likelihood ratio
- ⋆ Fisher's exact test

- Results

Ultimate goal

Provide a description of the linguistic behavior of **fixed expressions** in a computational language model

- *(g)een* ⟨ adjective ⟩ *figuur slaan* 'to make a ⟨ adjective ⟩ impression'
- *de aap uit de mouw komen* 'to reveal the truth'
- *iets tegen het licht houden* 'to investigate sth'
- *de plaat poetsen* 'depart unnoticed'
- *last hebben van*, *rekening houden met*

What are fixed expressions?

Recurrent word combinations

Ze slaan een goed figuur als ze mij contracteren .
wie hij in een campagne geen slecht figuur zou slaan .
et PSV-doelman Kralj een belachelijk figuur slaan en diens concurrent Waterreus
Honden laten je nooit een figuur slaan , integendeel , met een hond
de eerste maten zou ik geen slecht figuur slaan .
door uw uitingen een belachelijk figuur te slaan , en de angst door wat u ze
ken door paarden die een heel slecht figuur zouden slaan op een renbaan , rijden
en zeker met Christoph Daum geen gek figuur slaan , en danzou ik ze niet aanbeve
nwezigheid van de Koos Koets-achtige figuur is een uitkomst voor de tv-ploegen d
Nederland zou een plee-figuur slaan met Fortuyn naast Chirac , ald
stellen dat wij in Brussel een goed figuur slaan . "

Irregularities at various levels of linguistic description

Non-compositional meaning

(1) Bell **poetste de plaat** en La Primavera werd een hotel.
'Bell left unnoticed and La Primavera became a hotel.'

(2) Bell **poetste**/*maakte **de plaat** *schoon en ...
Bell cleaned/*made the plate *clean and ...

(3) **iets in bezit hebben** to own something

Rigid syntax and morphology

(4) Bell **poetste de** (*onopgemerkte) **plaat**(*je) en ...
Bell cleaned the *unnoticed plate(*little plate) and ...

(5) Ik geloof je niet, je **houdt** jezelf **voor** *__(de) gek__.
'I don't believe you, you consider yourself a fool.'

(6) Ik geloof je niet, je **houdt** jezelf **voor de** *__bange gek__.

Exceptions are numerous.

(7) We zitten **in elkaars/hun/jullie vaarwater**
'We work against each other/them/you.'

(8) Na de val van die andere grote ideologie (van het communisme)
'After the fall of the other big ideology (the communism),

zijn er in Oost-Europa weer **gevaarlijke** *apen uit de mouw gekomen*
dangerous truths were revealed in Eastern Europe.'

The first problem . . .

Using a corpus-based method, can we identify those expressions that qualify as fixed expressions?

A hybrid approach to identification

- The lexical fixedness between component words that bear a syntagmatic relationship is captured as a strong statistical dependence.

Nu zijn we bezig en dan [*houd*] *je je werk een keer* [*tegen het licht*]

Needed: association statistics to identify word combinations showing

- syntagmatic relationship between component words, and
- strong statistical dependence

Procedure

- **data preparation**
  - ⋆ annotate large corpus with linguistic information
- **dataset extraction**
  - ⋆ extraction of VERB PREPOSITIONAL PHRASE (PP) observations (eg. *houd [voor gek]*, *zit [in vaarwater]*)
  - ⋆ candidates represented VERB PREP NOUN triples
  - ⋆ determiners and adjectives ignored
- **identify effective association statistics**

Finding statistically dependent pairs

Measure the statistical dependence between the words inside candidate triples with pointwise mutual information, Pearson's χ^2, log-likelihood ratio, Fisher's exact test and salience.

Excerpt from the sample dataset. Observations

Total observations: 1167406

ben in Nederland	1241
heb te maken	910
ben in jaren	298
houd in gaten	125
zeg na afloop	96
houd tegen licht	29
ga op manier	26
kom uit mouw	8
zit op lip	6
houd van popmuziek	1

Frequency distribution. Contingency table

Categorical variables w_1 (word in position 1) and w_2 (word in position 2) have words as values. Assumed preposition + noun is one word.

To reduce the number of dimensions, assume each categorical variable is binomial (only two possible values). Cross-tabulate values for w_1 and w_2.

	w_2 in_gaten	$\neg\ w_2$ $\neg$ in_gaten	Row marginals
w_1 houd	O_{11} 125	O_{12} 7785	O_{1+} 7910
$\neg\ w_1$ $\neg$ houd	O_{21} 213	O_{22} 1159283	O_{2+} 1159496
Column marginals	O_{+1} =338	O_{+2}= 1167068	N = 1167406

Pointwise mutual information

$$I(w_1, w_2) = \log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$$

	w_2 in_gaten	$\neg\ w_2$ $\neg$ in_gaten	Row marginals
w_1 houd	O_{11} 125	O_{12} 7785	O_{1+} 7910
$\neg\ w_1$ $\neg$ houd	O_{21} 213	O_{22} 1159283	O_{2+} 1159496
Column marginals	O_{+1} =338	O_{+2}= 1167068	N = 1167406

MI example

$$I(houd, in\_gaten) = \log_2 \frac{P(houd, in\_gaten)}{P(houd)P(in\_gaten)} = \log_2 \frac{\frac{125}{1167406}}{\frac{7910}{1167406}\frac{338}{1167406}} =$$

$$\log_2 \frac{0.0001}{1.96e-06} = log_2 54.58 = 5.77$$

$$I(houd, met\_popmuziek) = \log_2 \frac{P(houd, met\_popmuziek)}{P(houd)P(met\_popmuziek)} =$$

$$\log_2 \frac{\frac{1}{1167406}}{\frac{7910}{1167406}\frac{1}{1167406}} = \log_2 \frac{8.56e-07}{7.33e-13} = log_2 \frac{1167406}{7910} = 20.15$$

Pearson's χ^2

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

	w_2 in_gaten	$\neg\ w_2$ $\neg$ in_gaten	Row marginals
w_1 houd	O_{11} 125	O_{12} 7785	O_{1+} 7910
$\neg\ w_1$ $\neg$ houd	O_{21} 213	O_{22} 1159283	O_{2+} 1159496
Column marginals	O_{+1} =338	O_{+2}= 1167068	N = 1167406

Limitations MI

- overestimates association score of rare events

χ^2 example

$$\chi^2 = (\frac{(125 - 2.29)^2}{2.29} + \frac{(7785 - 7907.7)^2}{7907.7} + \frac{(213 - 335.7)^2}{335.7} + \frac{(1159283 - 1159160)^2}{1159160}) = 6621.63$$

The χ^2 score for the low-frequency bigram $(houd, met\_popmuziek)$ is 146.58.

Log-likelihood ratio

$$G^2 = 2 \sum_{i,j} O_{ij} \log_2 \frac{O_{ij}}{E_{ij}}$$

Log-likelihood ratio example

$$G^2 = 2((125 \log_2 \frac{125}{2.29}) + (7785 \log_2 \frac{7785}{7907.71}) + (213 \log_2 \frac{213}{335.7}) + (1159283 \log_2 \frac{1159283}{1159160})) = 808.03 \quad (1)$$

Limitations χ^2 and G^2

Pedersen et al. (1996) show that G^2 and Pearson's χ^2 scores are unreliable when:

- the sample is not large enough,
- a cell's expected frequency is less or equal than 5

In my data: Pearson's χ^2 overestimates the significance of rare data. The corresponding G^2 and χ^2 scores attributed to the bigram $(houd, met\_popmuziek)$ are rather different; this suggests that the data is insufficient for the χ^2 score to be reliable.

Fisher's exact test

$$P(O_{11}) = \frac{\binom{O_{1+}}{O_{11}}\binom{O_{2+}}{O_{+1}-O_{11}}}{\binom{N}{O_{+1}}} \tag{2}$$

Candidates are ranked based on association scores

Frequency	VPN	Verb	Prep	Noun	chi-square
houd in gaten	222	14482	399605	674	11744.62
houd in stand	166	14482	399605	1674	8840.76
houd voor gezien	86	14482	153244	124	8660.64
heb in gaten	96	66433	399605	674	330.50
heb in hand	127	66433	399605	3133	308.85
ben in Nederland	2273	313023	399605	15515	365.04

Large association score reflects degree of lexical fixedness according to the specific statistic.

Results

- **Log-likelihood** and **Salience** test (a variant of pointwise mutual information) gives best results and are less sensitive to low-frequency data.

- Results are dependent on characteristics of your population (size, distribution of low-frequency vs. high-frequency candidates, noise added during pre-processing, etc.).