Two Multivariate Generalizations of Pointwise Mutual Information

Tim Van de Cruys

RCEAL, University of Cambridge

DISCO 2011
24 June 2011
Portland, OR

Introduction

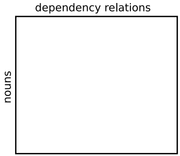
- pointwise mutual information: useful association measure in NLP
  - collocation extraction
  - weighting in vector space models
- restricted to two-way co-occurrences
- some NLP data can be tackled more naturally as multi-way co-occurrences

Two-way vs. three-way

- two way co-occurrence frequencies
- suitable for two-way problems
  - words $\times$ documents
  - nouns $\times$ dependency relations
- not suitable for *n*-way problems
  - words $\times$ documents $\times$ authors
  - verbs $\times$ subjects $\times$ direct objects

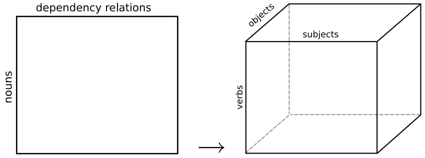
Two-way vs. three-way

- two way co-occurrence frequencies $\longrightarrow$ matrix
- suitable for two-way problems
  - words $\times$ documents
  - nouns $\times$ dependency relations
- not suitable for *n*-way problems
  - words $\times$ documents $\times$ authors
  - verbs $\times$ subjects $\times$ direct objects



Two-way vs. three-way

- two way co-occurrence frequencies $\longrightarrow$ matrix
- suitable for two-way problems
  - words $\times$ documents
  - nouns $\times$ dependency relations
- not suitable for *n*-way problems $\longrightarrow$ tensor
  - words $\times$ documents $\times$ authors
  - verbs $\times$ subjects $\times$ direct objects



Research question

- How to weight multi-way co-occurrences?
- two generalizations of pointwise mutual information, based on two different multivariate generalizations of mutual information

Mutual information

- measure of the amount of information that one random variable contains about another random variable
- reduction in the uncertainty of one random variable due to the knowledge of the other

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Pointwise mutual information

- measure of association that looks at particular instances of the two random variables X and Y
- measures difference between
  - the probability of their co-occurrence given their joint distribution
  - the probability of their co-occurrence given the marginal distributions of X and Y (thus assuming the two random variables are independent).

$$pmi(x, y) = \log \frac{p(x, y)}{p(x)p(y)}$$

Interaction information

- Interaction information (McGill, 1954) or co-information (Bell, 2003)
- based on the notion of conditional mutual information
- Conditional mutual information is the mutual information of two random variables conditioned on a third one

$$I(X;Y|Z)$$

$$= \sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)}$$

$$= \sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} p(x,y,z) \log \frac{p(z)p(x,y,z)}{p(x,z)p(y,z)}$$

Interaction information

- the interaction information: conditional mutual information subtracted by the standard mutual information
- Three-variable case:

$$\begin{aligned} I_1(X;Y;Z) &= I(X;Y|Z) - I(X;Y) \\ &= I(X;Z|Y) - I(X;Z) \\ &= I(Y;Z|X) - I(Y;Z) \end{aligned}$$

$$\begin{aligned} I_1(X;Y;Z) &= \sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} p(x,y,z) \log \frac{p(z)p(x,y,z)}{p(x,z)p(y,z)} \\ &- \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \end{aligned}$$

Specific interaction information

$$
\begin{aligned}
SI_1(x,y,z) &= \log \frac{p(x,y)}{p(x)p(y)} - \log \frac{p(z)p(x,y,z)}{p(x,z)p(y,z)} \\
&= \log \frac{p(x,y)p(y,z)p(x,z)}{p(x)p(y)p(z)p(x,y,z)}
\end{aligned}
$$

- can equally be defined for $n > 3$ variables

Total correlation

- total correlation (Watanabe, 1960) or multi-information (Studeny and Vejnarova, 1998)
- quantifies the amount of information that is shared among all random variables

$$I_2(X_1, X_2, \ldots, X_n)$$
$$= \sum_{\substack{x_1 \in X_1, \\ x_2 \in X_2, \\ \ldots \\ x_n \in X_n}} p(x_1, x_2, \ldots, x_n) \log \frac{p(x_1, x_2, \ldots, x_n)}{\prod_{i=1}^{n} p(x_i)}$$

Specific correlation

- correlation for specific instances of the random variables

$$SI_2(x_1, x_2, \ldots, x_n) = \log \frac{p(x_1, x_2, \ldots, x_n)}{\prod_{i=1}^{n} p(x_i)}$$

- For the case of three variables:

$$SI_2(x, y, z) = \log \frac{p(x, y, z)}{p(x)p(y)p(z)}$$

Extraction of svo-triples

- extraction of salient *subject verb object* triples (MWEs, fixed expressions)
- Experiment carried out for Dutch
- Twente Nieuws Corpus, parsed with Dutch dependency parser ALPINO
- *svo* triples with $f \geq 3$ extracted
- Construct tensor $\mathcal{T}$ of size $I \times J \times K$
  - $I = 1\text{K}$ verbs
  - $J = 10\text{K}$ subjects
  - $K = 10\text{K}$ objects

Extraction of svo-triples

- weight tensor using:
  - SI_1: specific interaction information

$$SI_1(x,y,z) = \log \frac{p(x,y)p(y,z)p(x,z)}{p(x)p(y)p(z)p(x,y,z)}$$

  - SI_2: specific correlation

$$SI_2(x,y,z) = \log \frac{p(x,y,z)}{p(x)p(y)p(z)}$$

Specific interaction information: example

- Top five *subject verb object* triples with highest *specific interaction information* score

subject	verb	object	SI_1
peiling 'poll'	*geef weer* 'represent'	*opinie* 'opinion'	18.20
helikopter 'helicopter'	*vuur af* 'fire'	*raket* 'rocket'	17.57
Man 'man'	*bijt* 'bite'	*hond* 'dog'	17.15
verwijt 'reproach'	*snijd* 'cut'	*hout* 'wood'	17.10
wal 'quay'	*keer* 'turn'	*schip* 'ship'	17.01

Specific correlation: example

- Top five *subject verb object* triples with highest *specific correlation* score

subject	verb	object	SI_2
verwijt	*snijd*	*hout*	8.05
'reproach'	'cut'	'wood'	
helikopter	*vuur af*	*raket*	7.75
'helicopter'	'fire'	'rocket'	
peiling	*geef weer*	*opinie*	7.64
'poll'	'represent'	'opinion'	
vlag	*dek*	*lading*	7.21
'flag'	'cover'	'load'	
argument	*snijd*	*hout*	7.17
'argument'	'cut'	'wood'	

Specific interaction information: example

- Top five combinations with highest *specific interaction information* scores for verb *speel* 'to play'

subject	verb	object	SI_1
orkest	*speel*	*symfonie*	11.65
'orchestra'	'play'	'symphony'	
leider	*speel*	*ruiten*	10.29
'leader'	'play'	'diamonds'	
leider	*speel*	*harten*	10.20
'leader'	'play'	'hearts'	
leider	*speel*	*schoppen*	10.01
'leader'	'play'	'spades'	
leider	*speel*	*klaveren*	9.89
'leader'	'play'	'clubs'	

Specific correlation: example

- Top five combinations with highest *specific correlation* scores for verb *speel* 'to play'

subject	verb	object	SI_2
nationaliteit 'nationality'	*speel* 'play'	*rol* 'role'	4.12
afkomst 'descent'	*speel* 'play'	*rol* 'role'	4.06
toeval 'coincidence'	*speel* 'play'	*rol* 'role'	4.04
motief 'motive'	*speel* 'play'	*rol* 'role'	4.04
afstand 'distance'	*speel* 'play'	*rol* 'role'	4.02

Quick evaluation

- Extract 100 triples with highest score for each method
- Evaluate triples according to salience
- Triple is considered salient if
  - made up of fixed (MWE) expression
  - fixed expression combined with prototypical argument (e.g. *argument snijd hout*)
- baseline: bare frequency tensor

measure	precision
baseline	.00
SI_1	.24
SI_2	.31

Rank correlation

- Kendall's τ_b to compare the rankings yielded by the two different methods
- Rank correlation calculated over all triples
- Correlation of $\tau_b = 0.21$
- results yielded by both methods – though correlated – differ to a significant extent

Conclusion

- not just one straightforward generalization of pointwise mutual information for the multivariate case
- two multivariate generalizations
  - *specific interaction information*
  - *specific correlation*
- useful for weighting multi-way co-occurrences in NLP tasks
  - extraction of salient *svo* triples
- Future work
  - More research into the exact nature of the dependencies that both measures capture
  - Proper quantitative evaluation on different multi-way co-occurrence tasks