Two Multivariate Generalizations of Pointwise Mutual Information

Tim Van de Cruys

RCEAL, University of Cambridge

DISCO 2011
24 June 2011
Portland, OR
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 × documents
    - nouns × dependency relations
  - not suitable for $n$-way problems
    - words × documents × authors
    - verbs × subjects × direct objects
Two-way vs. three-way

- Two way co-occurrence frequencies $\rightarrow$ 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 $\rightarrow$ matrix
- suitable for two-way problems
  - words $\times$ documents
  - nouns $\times$ dependency relations
- not suitable for $n$-way problems $\rightarrow$ 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:

\[ I_1(X; Y; Z) = I(X; Y|Z) - I(X; Y) \]
\[ = I(X; Z|Y) - I(X; Z) \]
\[ = I(Y; Z|X) - I(Y; Z) \]

\[ I_1(X; Y; Z) = \sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} p(x, y, z) \log \frac{p(z)p(x, y, z)}{p(x, z)p(y, z)} - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \]
Specific interaction information

\[ Sl_1(x, y, z) = \log \frac{p(x, y)}{p(x)p(y)} - \log \frac{p(z)p(x, y, z)}{p(x, z)p(y, z)} \]

\[ = \log \frac{p(x, y)p(y, z)p(x, z)}{p(x)p(y)p(z)p(x, y, z)} \]

- 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

\[
l_2(X_1, X_2, \ldots, X_n) = \sum_{x_1 \in X_1, x_2 \in X_2, \ldots, x_n \in X_n} p(x_1, x_2, \ldots, x_n) \log \frac{p(x_1, x_2, \ldots, x_n)}{\prod_{i=1}^{n} p(x_i)}
\]
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 $I \times J \times K$
  - $I = 1K$ verbs
  - $J = 10K$ subjects
  - $K = 10K$ objects
Extraction of svo-triples

- weight tensor using:
  - $Sl_1$: specific interaction information

\[
Sl_1(x, y, z) = \log \frac{p(x, y)p(y, z)p(x, z)}{p(x)p(y)p(z)p(x, y, z)}
\]

- $Sl_2$: specific correlation

\[
Sl_2(x, y, z) = \log \frac{p(x, y, z)}{p(x)p(y)p(z)}
\]
**Specific interaction information: example**

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

<table>
<thead>
<tr>
<th>subject</th>
<th>verb</th>
<th>object</th>
<th>$SI_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>peiling</td>
<td>geef weer</td>
<td>opinie</td>
<td>18.20</td>
</tr>
<tr>
<td>‘poll’</td>
<td>‘represent’</td>
<td>‘opinion’</td>
<td></td>
</tr>
<tr>
<td>helikopter</td>
<td>vuur af</td>
<td>raket</td>
<td>17.57</td>
</tr>
<tr>
<td>‘helicopter’</td>
<td>‘fire’</td>
<td>‘rocket’</td>
<td></td>
</tr>
<tr>
<td>Man</td>
<td>bijt</td>
<td>hond</td>
<td>17.15</td>
</tr>
<tr>
<td>‘man’</td>
<td>‘bite’</td>
<td>‘dog’</td>
<td></td>
</tr>
<tr>
<td>verwijt</td>
<td>snijd</td>
<td>hout</td>
<td>17.10</td>
</tr>
<tr>
<td>‘reproach’</td>
<td>‘cut’</td>
<td>‘wood’</td>
<td></td>
</tr>
<tr>
<td>wal</td>
<td>keer</td>
<td>schip</td>
<td>17.01</td>
</tr>
<tr>
<td>‘quay’</td>
<td>‘turn’</td>
<td>‘ship’</td>
<td></td>
</tr>
</tbody>
</table>
Specific correlation: example

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

<table>
<thead>
<tr>
<th>subject</th>
<th>verb</th>
<th>object</th>
<th>(SI_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>verwijt</td>
<td>snijd</td>
<td>hout</td>
<td>8.05</td>
</tr>
<tr>
<td>‘reproach’</td>
<td>‘cut’</td>
<td>‘wood’</td>
<td></td>
</tr>
<tr>
<td>helikopter</td>
<td>vuur af</td>
<td>raket</td>
<td>7.75</td>
</tr>
<tr>
<td>‘helicopter’</td>
<td>‘fire’</td>
<td>‘rocket’</td>
<td></td>
</tr>
<tr>
<td>peiling</td>
<td>geef weer</td>
<td>opinie</td>
<td>7.64</td>
</tr>
<tr>
<td>‘poll’</td>
<td>‘represent’</td>
<td>‘opinion’</td>
<td></td>
</tr>
<tr>
<td>vlag</td>
<td>dek</td>
<td>lading</td>
<td>7.21</td>
</tr>
<tr>
<td>‘flag’</td>
<td>‘cover’</td>
<td>‘load’</td>
<td></td>
</tr>
<tr>
<td>argument</td>
<td>snijd</td>
<td>hout</td>
<td>7.17</td>
</tr>
<tr>
<td>‘argument’</td>
<td>‘cut’</td>
<td>‘wood’</td>
<td></td>
</tr>
</tbody>
</table>
### Top five combinations with highest specific interaction information scores for verb speel ‘to play’

<table>
<thead>
<tr>
<th>subject</th>
<th>verb</th>
<th>object</th>
<th>$SI_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>orkest</td>
<td>speel</td>
<td>symfonie</td>
<td>11.65</td>
</tr>
<tr>
<td>‘orchestra’</td>
<td>‘play’</td>
<td>‘symphony’</td>
<td></td>
</tr>
<tr>
<td>leider</td>
<td>speel</td>
<td>ruiten</td>
<td>10.29</td>
</tr>
<tr>
<td>‘leader’</td>
<td>‘play’</td>
<td>‘diamonds’</td>
<td></td>
</tr>
<tr>
<td>leider</td>
<td>speel</td>
<td>harten</td>
<td>10.20</td>
</tr>
<tr>
<td>‘leader’</td>
<td>‘play’</td>
<td>‘hearts’</td>
<td></td>
</tr>
<tr>
<td>leider</td>
<td>speel</td>
<td>schoppen</td>
<td>10.01</td>
</tr>
<tr>
<td>‘leader’</td>
<td>‘play’</td>
<td>‘spades’</td>
<td></td>
</tr>
<tr>
<td>leider</td>
<td>speel</td>
<td>klaveren</td>
<td>9.89</td>
</tr>
<tr>
<td>‘leader’</td>
<td>‘play’</td>
<td>‘clubs’</td>
<td></td>
</tr>
</tbody>
</table>
Specific correlation: example

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

<table>
<thead>
<tr>
<th>subject</th>
<th>verb</th>
<th>object</th>
<th>$S_{I_2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>nationaliteit</em></td>
<td><em>speel</em></td>
<td><em>rol</em></td>
<td>4.12</td>
</tr>
<tr>
<td>‘nationality’</td>
<td>‘play’</td>
<td>‘role’</td>
<td></td>
</tr>
<tr>
<td><em>afkomst</em></td>
<td><em>speel</em></td>
<td><em>rol</em></td>
<td>4.06</td>
</tr>
<tr>
<td>‘descent’</td>
<td>‘play’</td>
<td>‘role’</td>
<td></td>
</tr>
<tr>
<td><em>toeval</em></td>
<td><em>speel</em></td>
<td><em>rol</em></td>
<td>4.04</td>
</tr>
<tr>
<td>‘coincidence’</td>
<td>‘play’</td>
<td>‘role’</td>
<td></td>
</tr>
<tr>
<td><em>motief</em></td>
<td><em>speel</em></td>
<td><em>rol</em></td>
<td>4.04</td>
</tr>
<tr>
<td>‘motive’</td>
<td>‘play’</td>
<td>‘role’</td>
<td></td>
</tr>
<tr>
<td><em>afstand</em></td>
<td><em>speel</em></td>
<td><em>rol</em></td>
<td>4.02</td>
</tr>
<tr>
<td>‘distance’</td>
<td>‘play’</td>
<td>‘role’</td>
<td></td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>measure</th>
<th>precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>.00</td>
</tr>
<tr>
<td>$SI_1$</td>
<td>.24</td>
</tr>
<tr>
<td>$SI_2$</td>
<td>.31</td>
</tr>
</tbody>
</table>
Rank correlation

- Kendall’s $\tau_b$ 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
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