Assessing morphological productivity via automated measures of semantic transparency

Marco Baroni
Università di Bologna
Bologna, Italy
baroni@sslmit.unibo.it

Stefano Vegnaduzzo
Conversay Corporation
Redmond WA, U.S.A.
svegnaduzzo@conversay.com

Explaining Productivity Workshop
February 27, 2003
Introduction

- Semantic transparency is likely to play important role in understanding morphological productivity.

- *Transparency → Productivity*: Only if an affix/morphological process occurs in a number of semantically transparent forms learners can discover semantic properties of affix/process, and use it to create/parse new forms.

- *Productivity → Transparency*: Productive processes are used to create nonce forms, which must be semantically transparent.
Introduction (continued)

• Relationship between productivity and semantic transparency is hard to study, because the latter is hard to compute (not to mention about the former...)

• Manual assessment of semantic transparency:
  – Resource-intensive;
  – Ratings at best hard-to-interpret, at worst circular.

• Automated semantic transparency measure would solve both problems.
Outline

• Automated measures of semantic transparency.

• Empirical test of measures on English words beginning with re-.

• From semantic transparency to productivity: a first try.

• What should we try next?
Measuring semantic transparency:
The basic idea

• Computational linguists have developed methods to measure semantic similarity among words.

• Degree of semantic transparency of complex form seen as degree of semantic similarity between complex form and its base.
Measuring semantic transparency:
The contextual approach

• Contextual approach to meaning.

• Cruse (1986, p.1):

  [T]he semantic properties of a lexical item are fully reflected in appropriate aspects of the relations it contracts with actual and potential contexts [...] [T]here are are good reasons for a principled limitation to linguistic contexts.
Measuring semantic transparency: 
Shared context and direct co-occurrence

- Two knowledge-poor interpretations of contextual approach:
  - Semantically related words will tend to occur in similar contexts.
  - Semantically related words will tend to occur near each other.
Shared context

• Cosine similarity (correlation) of normalized contextual vectors (Manning and Schütze 1999, Ch. 8):

\[
\cos(\vec{x}, \vec{y}) = \vec{x} \cdot \vec{y} = \sum_{i=1}^{n} x_i y_i
\]

• Compare patterns of co-occurrence of target words (in our case: base/derived form) with each word in corpus.

• Window of co-occurrence: 10 words.

• Minimum co-occurrence threshold: 2 occurrences.
Direct co-occurrence

- Basic intuition: related words will tend to co-occur more often than what we would predict based on their unigram frequencies.

- Several related Association Measures (Evert 2001) are based on comparison of empirical frequency of co-occurrence of two words and expected frequency of co-occurrence under assumption of independence: The larger the discrepancy, the more likely it is that the two words are not independent.
Direct co-occurrence (continued)

- We use:
  - **Mutual information** (Church and Hanks 1990):
    \[
    MI(w_1, w_2) = \log \frac{Pr(w_1, w_2)}{Pr(w_1)Pr(w_2)}
    \]
  - **Log-likelihood ratio** (Dunning 1993):
    \[
    -2 \log \lambda = 2 \sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}}
    \]
Direct co-occurrence *(continued)*

- What counts as a “co-occurrence” of two words?
  
  - Co-occurrence of adjacent directed n-grams: finds collocations, technical terms, etc.
  

- We are interested in the second notion of co-occurrence.
Direct co-occurrence (continued)

- Co-occurrence window: 150 words.
- Minimum distance: 3 words.
- Non-directionality: $a...b = b...a$. 
Empirical test of semantic transparency measures

• Focus on English prefix re-.


• 250 words with highest document frequency removed.

• Cosine similarity, AMs and token frequency (as control) computed for the 1211 pairs matching following conditions:
  
  – reSTEM and STEM both attested with fq > 1.

  – STEM’s length (in characters) ≥ 4.

• (Token fq is fq of form beginning with re-);
Test set construction

• Because of usual Zipfian reasons, semantic similarity scores have long tail of extremely low values.

• A problem for random sampling.

• Instead, test set constructed as follows:
  – Rank pairs on the basis of each measure.
  – Divide each ranked list into fourths.
  – For each ranked list, randomly sample 35 pairs from top fourth, 5 pairs from each of the following three fourths.
Human ratings

- 5 linguistically sophisticated English native speakers rated resulting set of 188 words for semantic transparency, on scale from 1 to 5.

- Pairwise Spearman correlations of judges’ ratings:
  - Min: 0.69
  - Avg: 0.77
  - Max: 0.87

- Judges’ avg rating compared with cosine, AMs, fq scores in a series of Spearman correlation analyses.
Semantic transparency results

<table>
<thead>
<tr>
<th>measure</th>
<th>corr 188</th>
<th>corr 50</th>
</tr>
</thead>
<tbody>
<tr>
<td>co</td>
<td>0.00</td>
<td>-0.18</td>
</tr>
<tr>
<td>mi</td>
<td>0.39</td>
<td>0.51</td>
</tr>
<tr>
<td>ll</td>
<td>0.13</td>
<td>-0.06</td>
</tr>
<tr>
<td>fq</td>
<td>-0.46</td>
<td>-0.55</td>
</tr>
</tbody>
</table>
Discussion

• Clearly, more work needed before replacing humans with machines!

• MI’s notorious tendency to favor low frequency pairs – good for morphology? (same result emerging from work on unsupervised morphological learning).

• MI and frequency not correlated (Spearman: -0.08): use in combination?

• Cosine negative correlation mystery... complementary distribution of related forms?
Moving on to productivity...

- Focus on be- de- en- in- mis- re- un- ("Baayen’s prefixes").

- Same corpus, same base/prefixed form extraction methods as above (resulting in 3750 candidate pairs).

- “Semantics”-based productivity measures: avg cosine similarity, avg MI, avg log-likelihood ratio (for set of pairs corresponding to each prefix).

- Other measures: type frequency, hapax frequency, hapax frequency / token frequency (Baayen’s $P$) [no hapax filtering, 5042 pairs].
(Informed) human ranking

• 4 morphologists, native speakers, asked to rank prefixes in order of productivity.

• Very similar responses

• Collective rank based on sums of individual rank scores:
  
  - un- (26) re- (25.5)
  
  - mis- (19.5) de- (17)
  
  - be- (8.5) en- (8.5) in- (7)
Semantics-based measures

- Cosine:

  \[ \text{re- un- } \star \text{in- mis- } \star \text{en- de- be-} \]

- Mutual Information:

  \[ \text{un- } \star \text{in- re- mis- } \star \text{en- de- be-} \]

- Log-Likelihood Ratio:

  \[ \text{un- re- in- be- } \star \text{mis- en- } \star \text{de-} \]
Type Fq, Hapaxes and $\mathcal{P}$

- Type Frequency:

  re- un- de- in- be- en- *mis-

- Number of hapaxes:

  un- re- de- in- be- *mis- en-

- $\mathcal{P}$:

  un- mis- *be- de- en- *re- in-
Discussion

• (Assuming that human morphologists are right) no perfect measure, no disastrous measure...

• ... and different measures have problems with different prefixes...

• ... but simple measures perform best!

• (Similar results (with slight improvement in MI) with stemmed version of same corpus.)
What’s next?

- Empirical testing on a larger scale and with more attention to detail.
- Work on transparency measures and their interpretation.
- Work on deriving productivity index from transparency measures.