Verb polysemy and frequency effects in thematic fit modeling

Clayton Greenberg  Vera Demberg  Asad Sayeed

Introduction

Human judgements of thematic fit (e.g. how plausible (“croquet”, “soccer”, “harpsichord”, “cheese”) are as patients of “play”) are likely sensitive to:

1) verb polysemy

Hypothesis: high polysemy → low thematic fit. Avoid confound: the most polysemous verbs are the most frequent.

2) sense frequency

WordNet orders SynSets based on their frequencies. Hypothesis: more frequent senses get higher scores.

3) role-filler (noun) frequency

Soccer is most frequent, croquet is declining, and harpsichord is oldest. Hypothesis: does not affect thematic fit.

Stimuli selection

McRae et al. (1997) and others obtained human judgements on a scale from 1 (lowest fit) to 7 (highest fit).

Their question: How common is it for croquet to be played?
Our question: Croquet is something that is played. Agree?

Verb selection

MONOSEMOUS: frequent in COCA, 1 SynSet in WordNet
POLYSEMOUS: matched for COCA freq., many SynSets

Role-filler selection

To find a good patient-filler, query COCA for: VERB [at*] [nn*]. Example: “whip the horse”
Find a much higher or lower frequency synonym. Example: “horse” (32384)

For POLYSEMOUS verbs, repeat 1 and 2 for second sense. Example: “whip” “cream” (19727) “frosting” (905)

Analysis of human judgements

For POLYSEMOUS verbs, bad patient-fillers were not as bad and good patient-fillers were not as good (**).

Noun frequency had a small effect on ratings of good patient-fillers (**), but not on ratings of bad patient-fillers ( ).

Participants rated good patient-fillers for the more frequent sense higher than those for the less frequent sense (**).

Noun frequency affected ratings on POLYSEMOUS verbs (.) less than ratings on MONOSEMOUS verbs (**).

Modeling results

<table>
<thead>
<tr>
<th></th>
<th>POLYSEMOUS</th>
<th>MONOSEMOUS</th>
<th>FILLERS</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centroid</td>
<td>0.405</td>
<td>0.655</td>
<td>0.313</td>
<td>0.464</td>
</tr>
<tr>
<td>OneBest</td>
<td>0.447</td>
<td>0.641</td>
<td>0.223</td>
<td>0.452</td>
</tr>
<tr>
<td>kClusters</td>
<td>0.432</td>
<td>0.669</td>
<td>0.304</td>
<td>0.479</td>
</tr>
</tbody>
</table>

▶ Spearman’s ρ between human judgements and Greenberg, Sayeed, & Demberg (NAACL, 2015) estimates.
▶ These confirm that using multiple prototypes (OneBest, kClusters) is more helpful for POLYSEMOUS verbs than MONOSEMOUS verbs, and that clustering (kClusters) is best for mixed datasets.

Conclusions

▶ We generated the first dataset of thematic fit judgements that systematically manipulates verb polysemy and role-filler frequency, avoiding confounds with verb unigram frequency.
▶ Modeling results show: multiple prototypes per verb-role help with polysemy, but sense frequency should still be addressed.
▶ These effects help characterize the nature of linguistic and conceptual material activated by verbs and their arguments.
▶ Distributional Memory models and datasets are available at: http://rollen.mncl.uni-saarland.de