Semi-automatic Construction of Cross-period Thesaurus

Chaya Liebeskind, Ido Dagan, Jonathan Schler
Bar Ilan University

Support search over historical texts of two types:
• Ancient texts
• Modern texts quoting ancient language (mixed)

Researchers are typically aware of modern language

Cross-period (Diachronic) Thesaurus

<table>
<thead>
<tr>
<th>Target term (modern)</th>
<th>Ancient related terms</th>
<th>Modern related terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car</td>
<td>wagon, carriage</td>
<td>auto, motor</td>
</tr>
<tr>
<td>Ship</td>
<td>Dreadnought, Bireme</td>
<td>boat, craft</td>
</tr>
<tr>
<td>Weapon</td>
<td>sword, naboot</td>
<td>gun, bomb</td>
</tr>
</tbody>
</table>

Our task:
• Support searches of cultural resources
• High-quality diachronic thesaurus construction
  ➢ Semi-automatic setting

Automatically generation of candidates

Statistical approaches for semantic relatedness identification
• First-order co-occurrence (fits corpus nature)
• Second-order distributional similarity
  ♦ mediocrity quality (corpus dependent)

Co-occurrence-based methods

Assumption
Words that occur frequently together are topically related

Common metrics
Dice coefficient, Pointwise Mutual Information, log-likelihood

Semi-automatic thesaurus construction (non-iterative)

Modern target term

Candidate Extraction from Corpus

wagon, carriage

Manual Selection

Related terms

Cross-period Thesaurus

Limited recall
⇒ Only documents consisting the modern target term are utilized (only mixed corpora, not ancient)

Solution
Increase number of documents in the statistical extraction
• Utilize ancient corpora

Apply query expansion to increase coverage

Modern target term

Candidate Extraction from Corpus

wagon, carriage

Manual Selection

Related terms

Cross-period Thesaurus

Case study: Cross-period Jewish thesaurus

The Responsa Corpus
• Questions posed to rabbis along their detailed answers
• Written over about a thousand years
• 76,760 articles
• Used for previous IR and NLP research

Ancient Corpora
➢ Responsa documents from the 11th century until the 19th century

Mixed Corpora
➢ Responsa documents from the 20th century until today

Jewish law literature

Evaluation: over 100 modern test target terms
Compare the iterative scheme to a baseline of a similar non-iterative method
• Compare for same number of judgements

<table>
<thead>
<tr>
<th>Method</th>
<th>Baseline</th>
<th>Iterative</th>
<th>Rel. ( \Delta )</th>
</tr>
</thead>
<tbody>
<tr>
<td># judgments (J)</td>
<td>14626</td>
<td>14462</td>
<td>−</td>
</tr>
<tr>
<td># related terms (RT)</td>
<td>892</td>
<td>1106</td>
<td>0.24</td>
</tr>
<tr>
<td>Average RT per target term</td>
<td>9</td>
<td>11</td>
<td>0.24</td>
</tr>
<tr>
<td>Precision-Productivity (RT/J)</td>
<td>0.06</td>
<td>0.08</td>
<td>0.33</td>
</tr>
<tr>
<td>Relative Recall (R)</td>
<td>0.78</td>
<td>0.99</td>
<td>0.27</td>
</tr>
</tbody>
</table>

Conclusions and future work

• A new task: cross-period thesaurus construction
• Semi-automatic iterative QE scheme
• Increasing thesaurus coverage, while optimizing the lexicographer manual effort

Future work:
• Adopt second-order distributional similarity methods