BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network

Roberto Navigli,
Simone Paolo Ponzetto
What is BabelNet

- ‘a very large, wide-coverage multilingual semantic network’
- Built from WordNet and Wikipedia
- 3 million concepts
- Mapping Accuracy of 82%
- Covers 52% of noun senses in WordNet
- Lemmas in 6 languages
  - 72.3% of synsets contain all 6
the plays and novels of Samuel Beckett...
...based on Shakespeare’s play Othello...
...dramatic media (plays, films, etc.)...

dramatic force in A. Miller’s play...
as the play opens the audience...
characters in the play take...

Machine Translation system

Wikipedia

WordNet

play

History of theatre

Musical theatre

Grand Guignol

Crime fiction

literature

dialog

stage direction

actor’s line

literary

derived-from

Babel Synset

play

play_

drama_

obra_

Bühnenwerk_

obra_

Theaterstück_

opera teatrale_

dramma_

pièce de théâtre_
WordNet

- Synset(s) for each word
- Labeled, Hierarchical relations Relations
- Glosses

\[ \{ \text{play}^1_N, \text{drama}^1_N, \text{dramatic play}^1_N \} \]

\text{play}^1_N \text{ is a dramatic composition}^1_N \\
\text{Shakespeare}^1_N \text{ instance-of } \text{dramatist}^1_N \\
\text{Stage direction}^1_N \text{ part-of } \text{play}^1_N
- WordNet
  - Synset(s) for each word
  - Labeled, Hierarchical relations Relations
  - Glosses

\{play^8_N, \text{child's } play^2_n\}
• Wikipedia
  • Page title
    • Optional label
  • Structured Relations
    • Redirect pages
    • disambiguation pages
    • internal links
    • inter-language links
    • categories

PLAY
PLAY (THEATRE)
Building BabelNet

1. Map Wikipedia to WordNet
2. Harvest multilingual lexicalizations from Wikipedia
3. Identify relations between synsets using WordNet and Wikipedia articles from all relevant languages
Mapping Wikipedia to WordNet

1. All available word senses and labeled relationships between them are pulled from WordNet
2. All encyclopedic entries and unspecified relationships between them are pulled from English Wikipedia
3. Combine them, and merge any intersecting senses
   1. If a sense exists in Wikipedia, but not in WordNet, the sense gets a null mapping
   2. If there is only one sense in Wikipedia and one sense in WordNet, create a one-to-one mapping
   3. For all remaining Wikipedia senses, map using the mapping algorithm:

\[
\arg\max_{s \in \text{Senses}_{WN}(w)} p(s|w) = \arg\max_{s} p(s,w) \\
p(s,w) = \frac{\text{score}(s,w)}{\sum_{s' \in \text{Senses}_{WN}(w), w' \in \text{Senses}_{Wiki}(w)} \text{score}(s',w')} 
\]
Mapping Wikipedia to WordNet

\[ p(s,w) = \frac{\text{score}(s,w)}{\sum_{s' \in \text{Senses}_{\text{WN}}(w)} \text{score}(s',w')} \]

**Bag-of-words**

\[ \text{score}(s,w) = |\text{Ctx}(s) \cap \text{Ctx}(w)| + 1 \]

• Simple and fast

**Graph-based**

\[ \text{score}(s,w) = \sum_{cw \in \text{Ctx}(w)} \sum_{s' \in \text{Senses}_{\text{WN}}(cw)} \sum_{p \in \text{paths}_{\text{WN}}(s,s')} e^{-(\text{length}(p)-1)} \]

• Exploits structural information available in WordNet and Wikipedia
• Provides mappings even when the intersection between WordNet and Wikipedia disambiguation contexts is empty.
Disambiguation Contexts

**WordNet**

- Synonymy
- Hypernymy/hyponymy
- Gloss (lemmatized, non-stop words)

**Wikipedia**

- Sense labels
- Links
- Redirections
- Categories
Translating Synsets

Synsets included
1. The word
2. The WordNet synset
3. Wikipedia redirections to the word
4. Other language Wikipedia articles
   1. Inter-language links
   2. Redirections to the inter-language links in the target language
5. Machine Translation using Google Translate
   1. WordNet contexts using SemCor contexts
   2. Wikipedia contexts from inter-wiki links to the page (BabelCor)
Translating Synsets

- Machine translation was not executed on named entities from Wikipedia
  - These were identified by titles containing at least 2 words with initial uppercase letters
  - 94% validation on a sample of 100 pages

- Contextless translation was performed on monosemous words from WordNet.
Finding Semantic Relations

- Wikipedia:
  - Harvesting relations from hyperlinks within all Wikipedias in the relevant languages
    - Ex. Play (theatre) links to Acting through the German Wikipedia, but not the English Wikipedia
  - Evaluate the frequency of co-occurrence of word1 and word2 ($f_{w,w'}$) in a context.

\[
\frac{2 \times f_{w,w'}}{f_w + f_{w'}}
\]
Finding Semantic Relations

- **WordNet:**
  - Gather lemmatized, non-stop word bags of words from:
    - Synonyms, gloss words, and directly linked synsets
  - Weighting relations using the Dice coefficient to determine their strength

\[
\frac{2 \times |S \cap S'|}{|S| + |S'|}
\]
Evaluation: Mapping

- Gold standard of 1,000 sample wikipages
  - 505 non-empty mappings
- Experienced annotator provided the correct WordNet sense for each page
  - 2nd annotator sense tagged subset of 200 pages to evaluate annotation accuracy
- Inter annotator agreement of 0.9
Evaluation: Mapping

- Bag-of-words:
  - $F_1 = 65.1$ (just gloss)
  - $F_1 = 75.0$ (using taxonomy and gloss)

- Graph-based:
  - $F_1 = 77.7$ (depth 2, using taxonomy and gloss)

- Baseline:
  - $F_1 = 33.5$ (most frequent sense)
  - $F_1 = 31.9$ (random selection)
Evaluation: Translation

- Automatic
  - Compared coverage of BabelNet with gold standard corpora
  - Catalan & Spanish: Multilingual Central Repository
  - French: Wordnet Libre du Français
  - German: GermaNet subset of EuroWordNet for German
  - Italian: MultiWordNet
Evaluation: Translation

\[ \text{SynsetCov}(B, F) = \frac{\sum_{S_F \in F} \delta(S_B, S_F)}{|\{S_F \in F\}|} \]

\[ \text{WordCov}(B, F) = \frac{\sum_{S_F \in F} \sum_{s_F \in S_F} \delta'(S_B, S_F)}{|\{s_F \in S_F : S_F \in F\}|} \]

\[ \text{WordExtraCov}(B, F) = \frac{\sum_{S_e \in \varepsilon \setminus F} \delta(S_B, S_e)}{|\{S_F \in F\}|} \]

\[ \text{WordExtraCov}(B, F) = \frac{\sum_{S_e \in \varepsilon \setminus F} \sum_{s_e \in S_e} \delta'(S_B, S_e)}{|\{s_F \in S_F : S_F \in F\}|} \]
### Table 6
Coverage against gold-standard wordnets (percentages).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Links</td>
<td>Transl.</td>
<td>Transl.</td>
<td>All</td>
<td>Links</td>
<td>Transl.</td>
<td>Transl.</td>
<td>All</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Catalan</td>
<td>20.3</td>
<td>46.9</td>
<td>25.0</td>
<td>64.0</td>
<td>25.2</td>
<td>54.1</td>
<td>29.6</td>
<td>73.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>French</td>
<td>70.0</td>
<td>69.6</td>
<td>16.3</td>
<td>86.0</td>
<td>72.4</td>
<td>79.6</td>
<td>19.4</td>
<td>92.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>German</td>
<td>39.6</td>
<td>42.6</td>
<td>21.0</td>
<td>57.6</td>
<td>50.7</td>
<td>58.2</td>
<td>28.6</td>
<td>73.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Italian</td>
<td>28.1</td>
<td>39.9</td>
<td>19.7</td>
<td>52.9</td>
<td>40.0</td>
<td>58.0</td>
<td>28.7</td>
<td>73.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td>34.4</td>
<td>47.9</td>
<td>25.2</td>
<td>66.4</td>
<td>40.7</td>
<td>56.1</td>
<td>30.0</td>
<td>76.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 7
Extra coverage against gold-standard wordnets (percentages).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Links</td>
<td>Transl.</td>
<td>Transl.</td>
<td>All</td>
<td>Links</td>
<td>Transl.</td>
<td>Transl.</td>
<td>All</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Catalan</td>
<td>100</td>
<td>204</td>
<td>71</td>
<td>340</td>
<td>35</td>
<td>105</td>
<td>42</td>
<td>142</td>
<td></td>
<td></td>
</tr>
<tr>
<td>French</td>
<td>255</td>
<td>223</td>
<td>92</td>
<td>514</td>
<td>63</td>
<td>102</td>
<td>67</td>
<td>159</td>
<td></td>
<td></td>
</tr>
<tr>
<td>German</td>
<td>1349</td>
<td>940</td>
<td>367</td>
<td>2298</td>
<td>506</td>
<td>668</td>
<td>303</td>
<td>902</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Italian</td>
<td>160</td>
<td>234</td>
<td>83</td>
<td>419</td>
<td>87</td>
<td>153</td>
<td>68</td>
<td>213</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td>214</td>
<td>158</td>
<td>56</td>
<td>384</td>
<td>48</td>
<td>74</td>
<td>30</td>
<td>102</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Evaluation: Manual Translation

- Manual Evaluation
  - Evaluates the novel synsets
  - Random set of 600 Babel synsets
  - Manually evaluated by expert annotators

  - Wordnet only synsets: 72%
  - Wikipedia only synsets: 95%
  - WordNet and Wikipedia synsets: 80%
How this relates to our project

- It really doesn’t anymore
- We’re now working on topic identification and classification of journal articles, using attributes derived from the article context.
- BabelNet can be used for topic identification
  - … so they’re kind of related