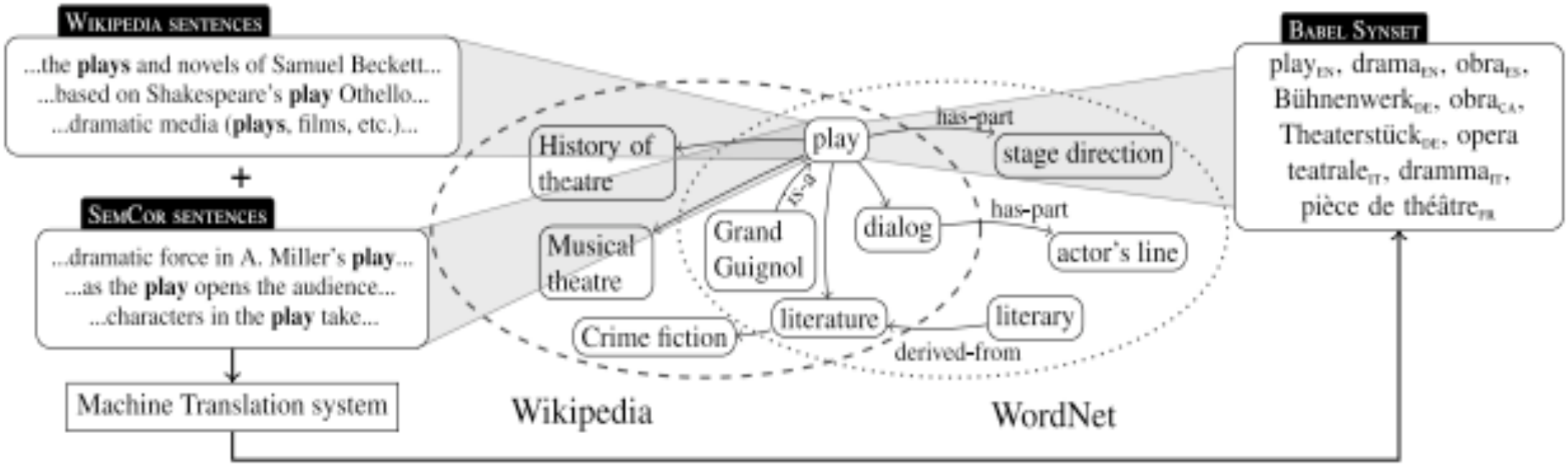
BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network

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


WIKIPEDIA SENTENCES
...the plays and novels of Samuel Beckett...
...based on Shakespeare's play Othello...
...dramatic media (plays, films, etc.)...
+
SEMCOR SENTENCES
...dramatic force in A. Miller's play...
...as the play opens the audience...
...characters in the play take...
Machine Translation system
BABEL SYNSET
play_EN, drama_EN, obra_ES, Bühnenwerk_DE, obra_CA, Theaterstück_DE, opera teatrale_IT, dramma_IT, pièce de théâtre_FR
History of theatre
play
has-part
stage direction
Musical theatre
Grand Guignol
is-a
dialog
has-part
actor's line
literature
literary
Crime fiction
derived-from
Wikipedia
WordNet

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

{play$^{1}_{n}$, drama$^{1}_{N}$, dramatic play$^{1}_{N}$}

play$^{1}_{N}$ *is a* dramatic composition$^{1}_{N}$

Shakespeare$^{1}_{N}$ *instance-of* dramatist$^{1}_{N}$

Stage direction$^{1}_{N}$ *part-of* play$^{1}_{N}$

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

{play$^{8}_{N}$, 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:

$$\arg\max_{s \in Senses_{WN}(w)} p(s|w) = \arg\max_{s} p(s,w)$$

$$p(s,w) = \frac{score(s,w)}{\sum_{\substack{s' \in Senses_{WN}(w), \\ w' \in Senses_{Wiki}(w)}} score(s',w')}$$

Mapping Wikipedia to WordNet

$$p(s,w) = \frac{score(s,w)}{\sum_{\substack{s' \in Senses_{WN}(w), \\ w' \in Senses_{Wiki}(w)}} score(s',w')}$$

Bag-of-words

$$score(s,w) = \left| Ctx(s) \cap Ctx(w) \right| + 1$$

- Simple and fast

Graph-based

$$score(s,w) = \sum_{cw \in Ctx(w)} \sum_{s' \in Senses_{WN}(cw)} \sum_{p \in paths_{WN}(s,s')} e^{-(length(p)-1)}$$

- 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

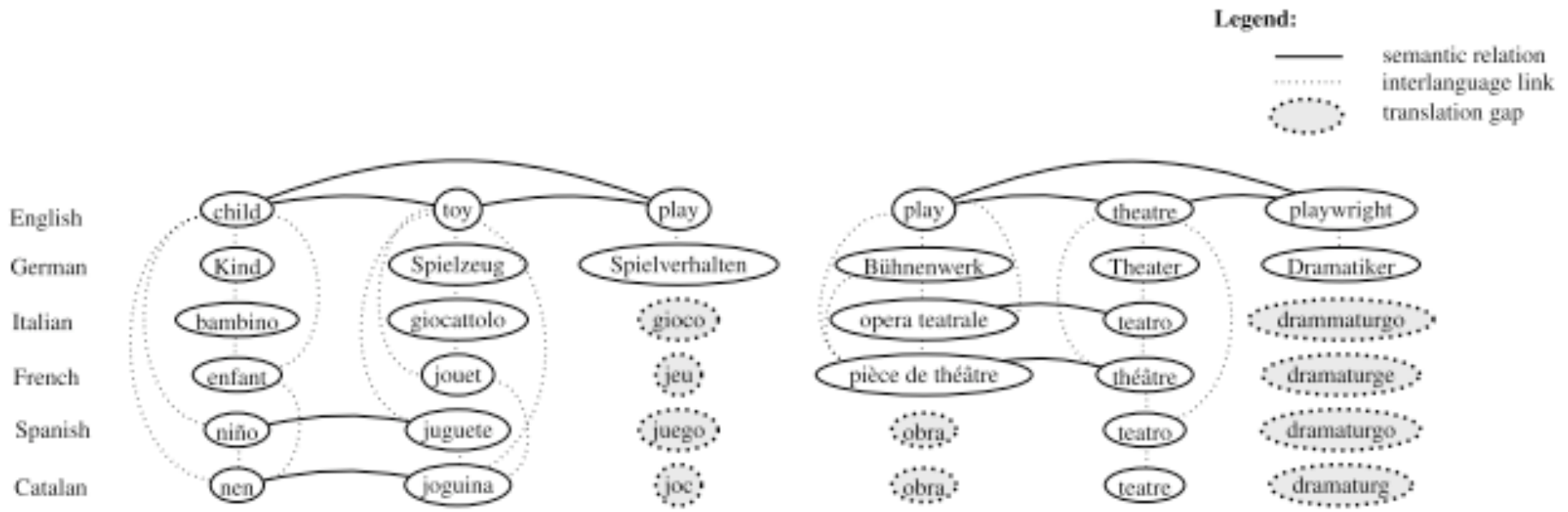
Translating Synsets

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.


Legend:
semantic relation
interlanguage link
translation gap
English
German
Italian
French
Spanish
Catalan
child
toy
play
Kind
Spielzeug
Spielverhalten
bambino
giocattolo
gioco
enfant
jouet
jeu
niño
juguete
juego
nen
joguina
joc
play
theatre
playwright
Bühnenwerk
Theater
Dramatiker
opera teatrale
teatro
drammaturgo
pièce de théâtre
théâtre
dramaturge
obra
teatro
dramaturgo
obra
teatre
dramaturg

Finding Semantic Relations

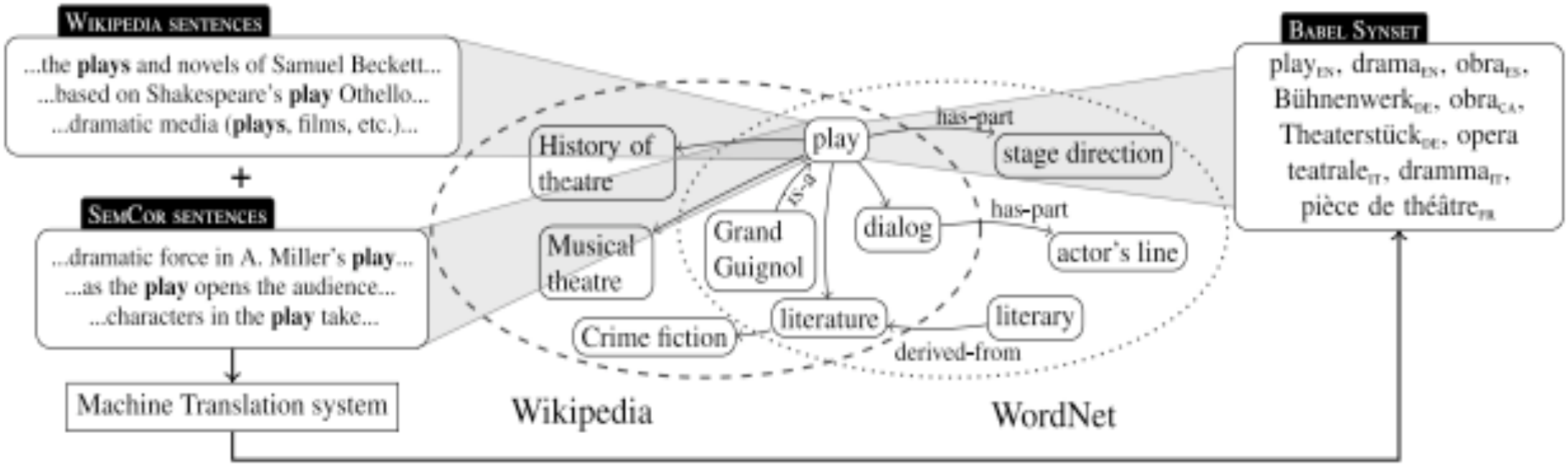
- Wikipedia:
  - Harvesting relations from hyperlinks within *all* Wikipedias in the relevant languages
    - Ex. Play (theatre) links to Acting through the German Wikipedia, but not the English Wikipedia
  - Evaluate the frequency of co-occurrence of word1 and word2 ($f_{w.w'}$) in a context.

$$\frac{2 \times f_{w.w'}}{f_w + f_{w'}}$$

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

$$\frac{2 \times |S \cap S'|}{|S| + |S'|}$$


WIKIPEDIA SENTENCES
...the plays and novels of Samuel Beckett...
...based on Shakespeare's play Othello...
...dramatic media (plays, films, etc.)...
+
SEMCOR SENTENCES
...dramatic force in A. Miller's play...
...as the play opens the audience...
...characters in the play take...
Machine Translation system
BABEL SYNSET
play_EN, drama_EN, obra_ES, Bühnenwerk_DE, obra_CA, Theaterstück_DE, opera teatrale_IT, dramma_IT, pièce de théâtre_FR
History of theatre
play
has-part
stage direction
Musical theatre
Grand Guignol
is-a
dialog
has-part
actor's line
literature
literary
Crime fiction
derived-from
Wikipedia
WordNet

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)
  - F1 _ 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

$$SynsetCov(B,F) = \frac{\sum_{S_F F} \delta(S_B, S_F)}{\left|\{S_F \in F\}\right|}$$

$$WordCov(B,F) = \frac{\sum_{S_F \in F} \sum_{s_F \in S_F} \delta'(S_B, S_F)}{\left|\{s_F \in S_F : S_F \in F\}\right|}$$

$$WordExtraCov(B,F) = \frac{\sum_{S_\varepsilon \in \varepsilon \setminus F} \delta(S_B, S_\varepsilon)}{\left|\{S_F \in F\}\right|}$$

$$WordExtraCov(B,F) = \frac{\sum_{S_\varepsilon \in \varepsilon \setminus F} \sum_{s_\varepsilon \in S_\varepsilon} \delta'(S_B, S_\varepsilon)}{\left|\{s_F \in S_F : S_F \in F\}\right|}$$

Evaluation: Automatic Translation

Table 6
Coverage against gold-standard wordnets (percentages).

	WordCov (Senses)				SynsetCov (Synsets)			
Resource	Wiki		WordNet	BabelNet	Wiki		WordNet	BabelNet
Method	Links	Transl.	Transl.	All	Links	Transl.	Transl.	All
Catalan	20.3	46.9	25.0	**64.0**	25.2	54.1	29.6	**73.3**
French	70.0	69.6	16.3	**86.0**	72.4	79.6	19.4	**92.9**
German	39.6	42.6	21.0	**57.6**	50.7	58.2	28.6	**73.4**
Italian	28.1	39.9	19.7	**52.9**	40.0	58.0	28.7	**73.7**
Spanish	34.4	47.9	25.2	**66.4**	40.7	56.1	30.0	**76.6**

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

	WordExtraCov (Senses)				SynsetExtraCov (Synsets)			
Resource	Wiki		WordNet	BabelNet	Wiki		WordNet	D..BabelNet
Method	Links	Transl.	Transl.	All	Links	Transl.	Transl.	All
Catalan	100	204	71	**340**	35	105	42	**142**
French	255	223	92	**514**	63	102	67	**159**
German	1349	940	367	**2298**	506	668	303	**902**
Italian	160	234	83	**419**	87	153	68	**213**
Spanish	214	158	56	**384**	48	74	30	**102**

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