

Semantic Indexing of Multilingual Corpora and its Application on the History Domain

Alessandro Raganato, Jose Camacho-Collados, Antonio Raganato and Yunseo Joung

Semantic Resources

BabelNet
-Multilingual Knowledge Base-

The screenshot shows the BabelNet search interface. At the top, there is a search bar with 'Edward' entered, a language dropdown set to 'ENGLISH', and a 'TRANSLATE' button. Below the search bar, there are filters for 'All', 'Concepts', and 'Named Entities'. The results are categorized under 'Noun' and list several instances of 'Edward' with their respective descriptions and multilingual labels. For example, 'Duke of Windsor, Edward VIII, Edward' is described as 'King of England and Ireland in 1936; his marriage to Wallis Warfield Simpson created a constitutional crisis leading to his abdication (1894-1972)'. Other instances include 'Edward VII, Albert Edward, Edward' and 'Edward, Edward VI, Edward VI of England'.

Babelfy

-Multilingual Disambiguation and Entity Linking-

The screenshot shows the Babelfy interface for the entity 'Napoleon Bonaparte'. It displays the sentence: 'Napoleon Bonaparte was a French military and political leader during the French Revolution'. Below the sentence, there are four boxes representing different senses of the entity: 'French' (Of or pertaining to France or the people of France), 'military' (Of or relating to the study of the principles of warfare), 'political leader' (A person active in party politics), and 'French Revolution' (The revolution in France against the Bourbons; 1789-1799). Each box includes a small image and a brief description. The Babelfy logo is also visible.

Semantic Indexing of Multilingual Corpora

Based on the pre-disambiguation of a corpus (Babelfy)

Search by **concept/entity**. Two main advantages over keyword searches:

- Unambiguous search
- Search in different languages

The screenshot shows the 'Input synset' and 'Retrieved texts' for the concept 'Edward IV of England'. The input synset includes the concept name in multiple languages (English, Spanish, Italian, Japanese, Chinese, Korean) and its category (NOUN, Named Entity). The retrieved texts show snippets from various languages, such as Korean, Chinese, Japanese, Spanish, and Italian, all referring to Edward IV of England. For example, the Spanish text reads: '... el 14 de mayo de 1471 fue asesinado y el trono volvió a manos de Eduardo IV nuevamente tras vencer a Warwick y Margarita d'Anjou en las batallas de Barnet y Tewkesbury [...]'.

XML format

```
<dataset language="EN" title="GEN">
<paragraph id="p.1">
<text>
In the beginning God created the heaven and the earth.
And the earth was without form, and void; and darkness was upon the face of the deep.
And the Spirit of God moved upon the face of the waters.
</text>
</text>
<annotations>
<annotation source="MCS" anchor="beginning" bfScore="--" coherenceScore="--">bn:00009632n</annotation>
<annotation source="BABELFY" anchor="God" bfScore="0.7620" coherenceScore="0.7913">bn:00040878n</annotation>
<annotation source="MCS" anchor="created" bfScore="--" coherenceScore="--">bn:00086006v</annotation>
<annotation source="BABELFY" anchor="earth" bfScore="0.8485" coherenceScore="0.8079">bn:00029424n</annotation>
...
</annotations>
</paragraph>
...

```

Disambiguation quality

Evaluation corpus: Two chapters of the Bible

Baseline: MCS (Most Common Sense)

Results (Precision):

		English	Spanish
All	Our	68.8	58.8
	MCS	51.1	44.0
Nouns	Our	74.2	63.4
	MCS	58.7	47.8

Cross-lingual text retrieval

Task: Given a paragraph/text as input, retrieve its most similar texts (in any language)

Method:

- Texts are represented as a set of unambiguous concepts/entities
- Similarity measure: Jaccard coefficient for sets

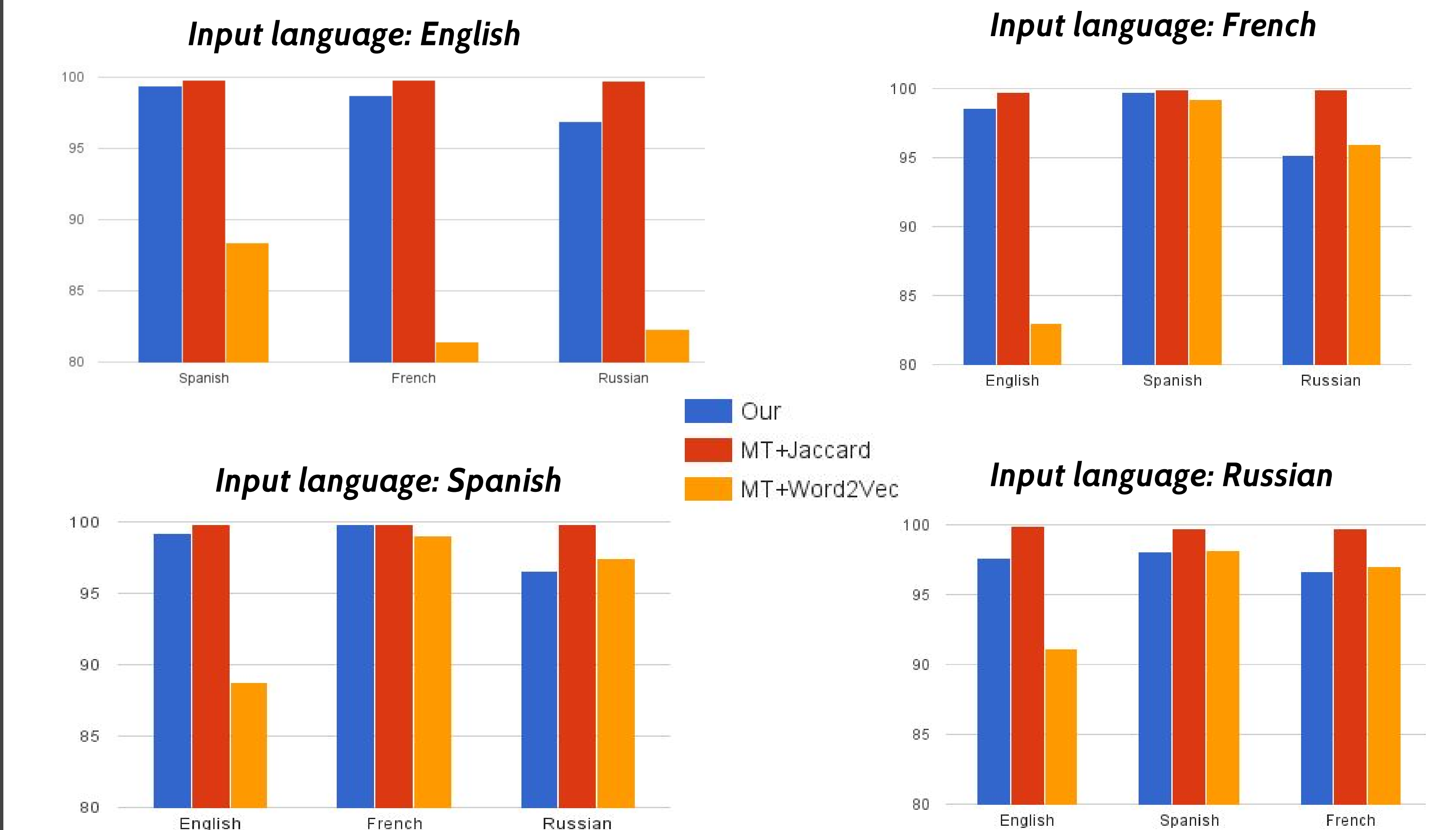
Experiments

Evaluation corpus: The Bible

Languages: English, French, Russian and Spanish

Task: Given an input chapter in the input language, retrieve the same chapter in the output language.

Baselines: Jaccard similarity after Machine Translation (MT+Jaccard) and Cosine similarity of the average of word embeddings after Machine Translation (MT+Word2Vec)



Release: Data and Interface

<http://wwwusers.di.uniroma1.it/~raganato/semantic-indexing/>

References

- Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217-250.
- Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231-244.