DIPARTIMENTO DI INFORMATICA
DIPARTIMENTO DI SCIENZE POLITICHE
SAPIENZA UNIVERSITÀ DI ROMA

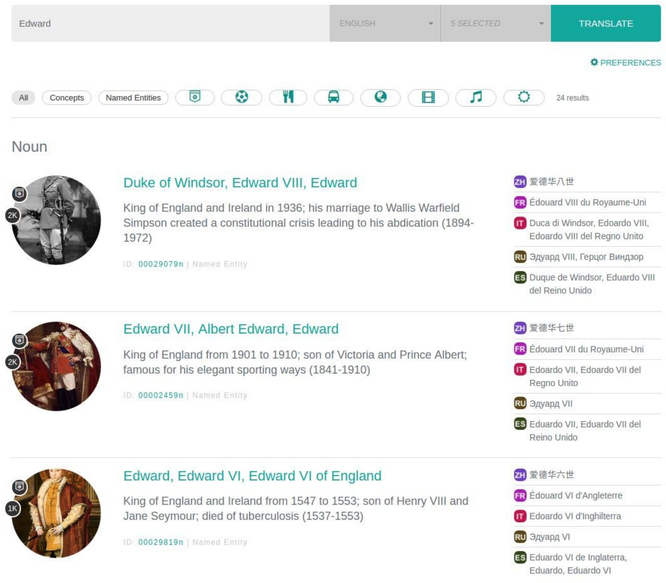
The authors gratefully acknowledge the support of the Sapienza Research Grant Avvio alla Ricerca 2015 No. 56

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

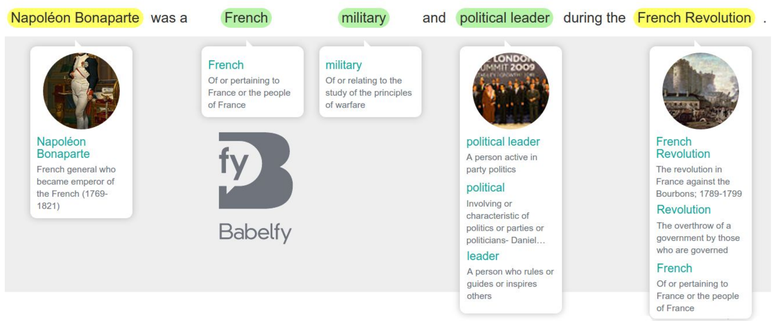
Alessandro Raganato, Jose Camacho-Collados, Antonio Raganato and Yunseo Joung

Semantic Resources

BabelNet
-Multilingual Knowledge Base-



Babelfy
-Multilingual Disambiguation and Entity Linking-

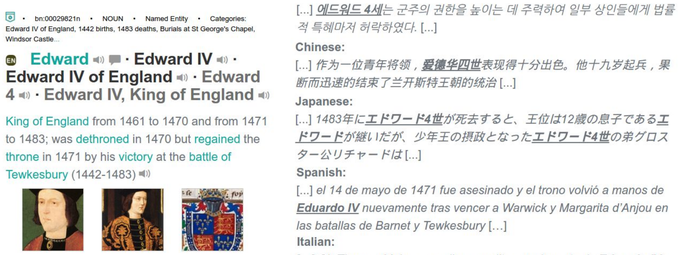


Semantic Indexing of Multilingual Corpora

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**



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>
    <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:00086008v</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

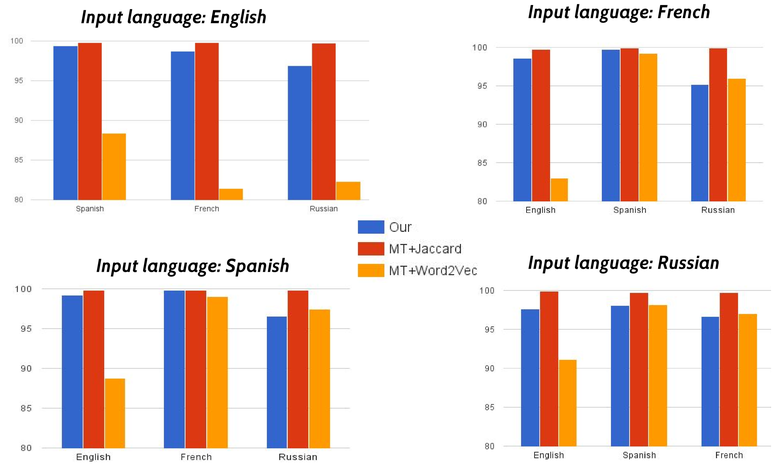
Experiments

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.