Automatic Identification and Disambiguation of Concepts and Named Entities in the Multilingual Wikipedia

Federico Scozzafava, Alessandro Raganato and Roberto Navigli

Dipartimento di Informatica

Sapienza Università di Roma

Linguistic Computing Laboratory

http://lcl.uniroma1.it

ERC Starting Grant MultiJEDI No. 259234

Babelfied Wikipedia: An annotated multilingual corpus

- **Goal**: Create a multilingual annotated corpus
  - With both word senses (i.e. concepts) and entities
- Calculate some statistics on that
- Automatically!

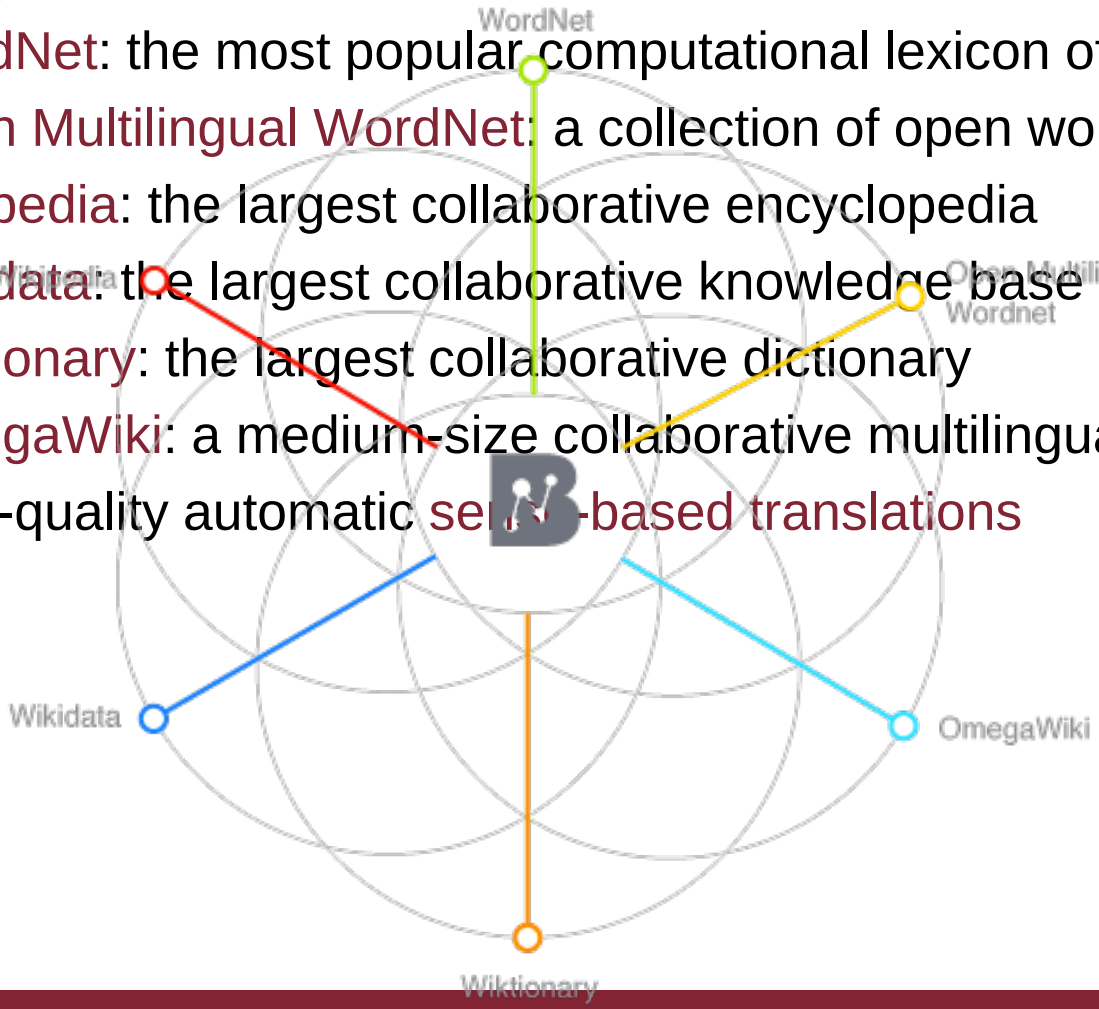
- *The annotated dataset is available for download in text and RDF/NIF format at http://lcl.uniroma1.it/babelfied-wikipedia/*

How are we going to do that?

BabelNet 3.5 (http://babelnet.org)

What is BabelNet

- A merger of resources of different kinds:
  - WordNet: the most popular computational lexicon of English
  - Open Multilingual WordNet: a collection of open wordnets
  - Wikipedia: the largest collaborative encyclopedia
  - Wikidata: the largest collaborative knowledge base
  - Wiktionary: the largest collaborative dictionary
  - OmegaWiki: a medium-size collaborative multilingual dictionary
  - High-quality automatic semantic-based translations



What is BabelNet

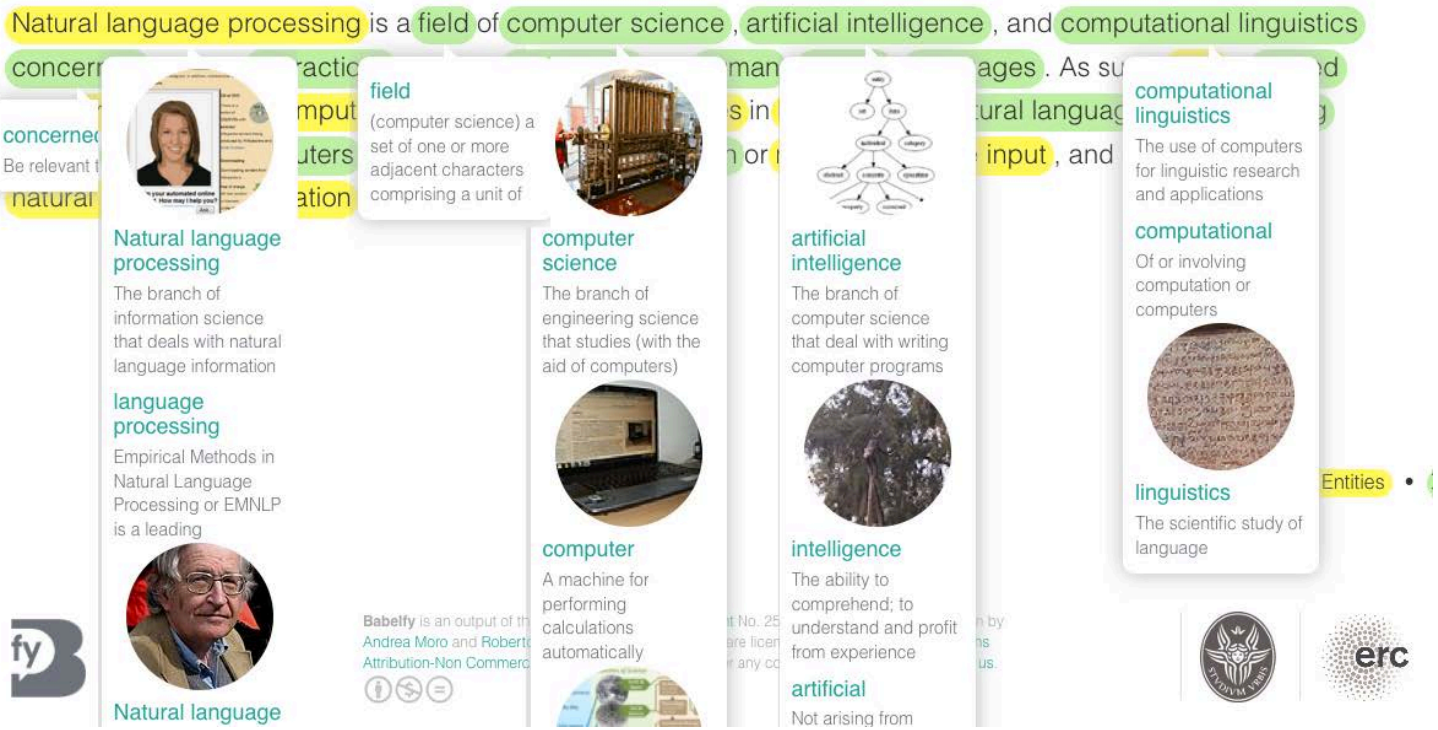
- Multilinguality: the same concept is expressed in tens of languages
- Coverage: 272 languages and 14 million entries!
- Concepts and named entities together: dictionary and encyclopedic knowledge is semantically interconnected
- Full-fledged taxonomy: is-a relations are available for both concepts and named entities (Wikipedia Bitaxonomy)
- [NEW] Semantic relations: semantic network structure with labeled relations from Wikidata and infoboxes
- [NEW] Domain labels for millions of synsets

How are we going to do that?

Babelfy (http://babelfy.org)

Babelfy: A Joint approach to WSD and Entity Linking [Moro et al., TACL 2014]

- **Babelfy** is a state-of-the-art unified graph-based approach to **Entity Linking** and **Word Sense Disambiguation** based on **BabelNet**



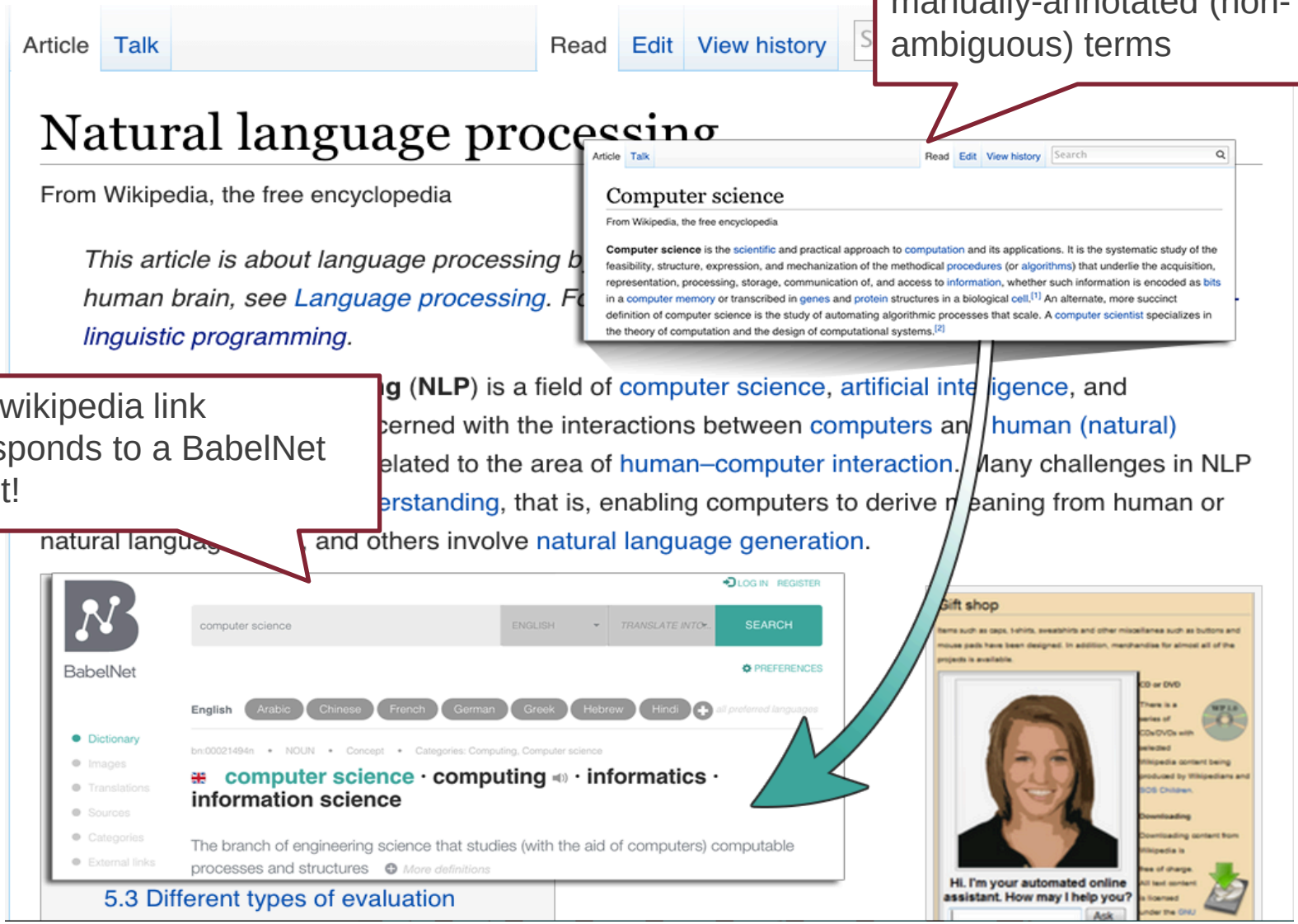
Disambiguating Wikipedia

- We applied **Babelfy 1.0** to the English and Italian editions of Wikipedia, disambiguating most of the content words.
- We used the **user-provided hyperlinks** to **improve** the quality of our automatic annotation.

Disambiguating Wikipedia

Wikipedia links provide manually-annotated (non-ambiguous) terms

Each wikipedia link corresponds to a BabelNet synset!



Statistics

For statistics and evaluation come to the social session!

Automatic Identification and Disambiguation of Concepts and Named Entities in the Multilingual Wikipedia

SAPIENZA
UNIVERSITÀ DI ROMA

Federico Scozzafava, **Alessandro Raganato**,
Andrea Moro and **Roberto Navigli**

{raganato,moro,navigli}@di.uniroma1.it

federico.scozzafava@gmail.com