

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

Babelified 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/babelified-wikipedia/>*

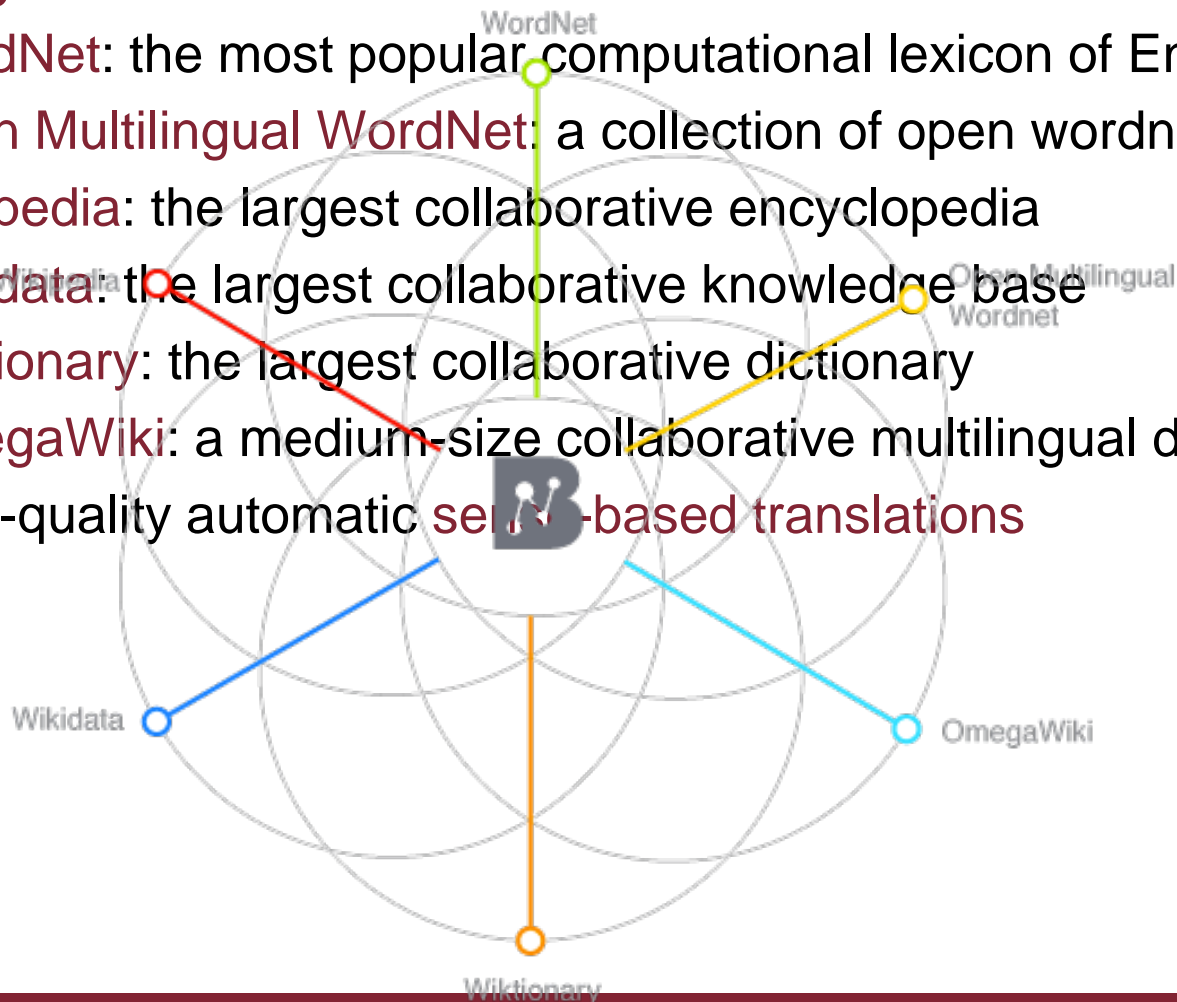
How are we going to do that?

BabelNet 3.5 (<http://babelnet.org>)



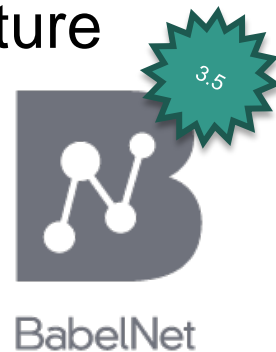
What is BabelNet

- 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 **sense-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**

The screenshot displays the Babelfy interface for the word "Natural language processing". It shows a list of related terms and their definitions, each with a representative image. The terms include:

- Natural language processing**: The branch of information science that deals with natural language information. Image: A woman's portrait.
- language processing**: Empirical Methods in Natural Language Processing or EMNLP is a leading. Image: A man's portrait.
- computer science**: The branch of engineering science that studies (with the aid of computers). Image: A server room.
- computer**: A machine for performing calculations automatically. Image: A laptop.
- artificial intelligence**: The branch of computer science that deal with writing computer programs. Image: A tree.
- computational linguistics**: The use of computers for linguistic research and applications. Image: A circular image of text.
- linguistics**: The scientific study of language. Image: A circular image of text.

At the bottom, there are logos for Babelfy, the University of Turin, and ERC.

Disambiguating Wikipedia

- We applied **Babely 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.



WIKIPEDIA
The Free Encyclopedia

Disambiguating Wikipedia

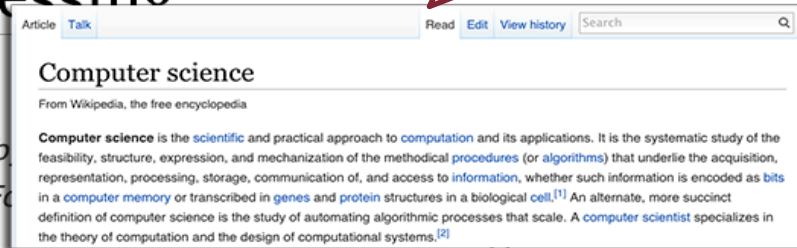
Wikipedia links provide manually-annotated (non-ambiguous) terms



Natural language processing

From Wikipedia, the free encyclopedia

This article is about language processing by the human brain, see [Language processing](#). For linguistic programming, see [linguistic programming](#).



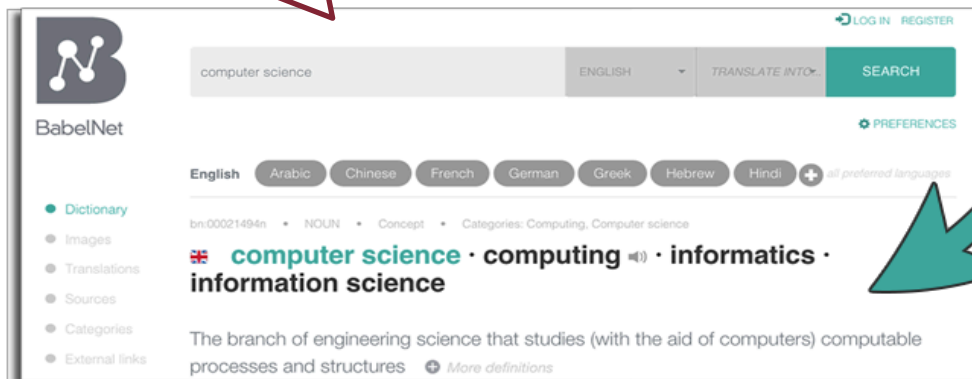
Computer science

From Wikipedia, the free encyclopedia

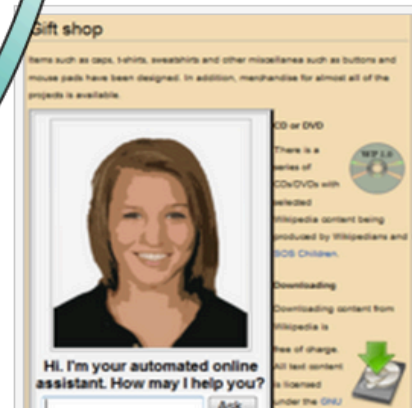
Computer science is the scientific and practical approach to [computation](#) and its applications. It is the systematic study of the feasibility, structure, expression, and mechanization of the methodical [procedures](#) (or [algorithms](#)) that underlie the acquisition, representation, processing, storage, communication of, and access to [information](#), whether such information is encoded as [bits](#) in a [computer memory](#) or transcribed in [genes](#) and [protein](#) structures in a biological [cell](#).^[1] An alternate, more succinct definition of computer science is the study of automating algorithmic processes that scale. A [computer scientist](#) specializes in the theory of computation and the design of computational systems.^[2]

Each wikipedia link corresponds to a BabelNet synset!

g (NLP) is a field of [computer science](#), [artificial intelligence](#), and [natural language processing](#) concerned with the interactions between [computers](#) and [human \(natural\) language](#) related to the area of [human–computer interaction](#). Many challenges in NLP are [natural language understanding](#), that is, enabling computers to derive meaning from human or machine [natural language](#) and others involve [natural language generation](#).



5.3 Different types of evaluation



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