(Digital) Goodies from the ERC Wishing Well: BabelNet, Babelfy, Video Games with a Purpose and the Wikipedia Bitaxonomy

Roberto Navigli
Dipartimento di Informatica
Sapienza Università di Roma
Viale Regina Elena, 295 – 00166 Roma Italy
navigli@di.uniroma1.it

Abstract

Multilinguality is a key feature of today's Web, and it is this feature that we leverage and exploit in our research work at the Sapienza University of Rome's Linguistic Computing Laboratory, which I am going to overview and showcase in this talk.

I will start by presenting **BabelNet 2.5** (Navigli and Ponzetto, 2012), available at http://babelnet.org, a very large multilingual encyclopedic dictionary and semantic network, which covers 50 languages and provides both lexicographic and encyclopedic knowledge for all the open-class parts of speech, thanks to the seamless integration of WordNet, Wikipedia, Wiktionary, OmegaWiki, Wikidata and the Open Multilingual WordNet. In order to construct the BabelNet network, we extract at different stages: from WordNet, all available word senses (as *concepts*) and all the lexical and semantic pointers between synsets (as *relations*); from Wikipedia, all the Wikipages (i.e., Wikipages, as *concepts*) and semantically unspecified *relations* from their hyperlinks. WordNet and Wikipedia overlap both in terms of concepts and relations: this overlap makes the merging between the two resources possible, enabling the creation of a *unified knowledge resource*. In order to enable multilinguality, we collect the lexical realizations of the available concepts in different languages. Finally, we connect the multilingual Babel synsets by establishing semantic relations between them.

Next, I will present **Babelfy** (Moro et al., 2014), available at http://babelfy.org, a unified approach that leverages BabelNet to perform Word Sense Disambiguation (WSD) and Entity Linking in arbitrary languages, with performance on both tasks on a par with, or surpassing, those of task-specific state-of-the-art supervised systems. Babelfy works in three steps: first, given a lexicalized semantic network, we associate with each vertex, i.e., either concept or named entity, a semantic signature, that is, a set of related vertices. This is a preliminary step which needs to be performed only once, independently of the input text. Second, given a text, we extract all the linkable fragments from this text and, for each of them, list the possible meanings according to the semantic network. Third, we create a graph-based semantic interpretation of the whole text by linking the candidate meanings of the extracted fragments using the previously-computed semantic signatures. We then extract a dense subgraph of this representation and select the best candidate meaning for each fragment. Our experiments show state-of-the-art performances on both WSD and EL on 6 different datasets, including a multilingual setting.

In the third part of the talk I will present two novel approaches to large-scale knowledge acquisition and validation developed in my lab. I will first introduce **video games with a purpose** (Vannella et al., 2014), a novel, powerful paradigm for the large scale acquisition and validation of knowledge and data (http://knowledgeforge.org). We demonstrate that converting games with a purpose into more traditional video games provides a fun component that motivates players to annotate for free, thereby significantly lowering annotation costs below that of crowdsourcing. Moreover, we show that video games with a purpose produce higher-quality annotations than crowdsourcing.

Zock/Rapp/Huang (eds.): Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon, pages 75–76,
Dublin, Ireland, August 23, 2014.

Then I will introduce the **Wikipedia Bitaxonomy** (Flati et al., 2014, WiBi), available at `http://wibitaxonomy.org` and now integrated into BabelNet. WiBi is the largest and most accurate currently available taxonomy of Wikipedia pages and taxonomy of categories, aligned to each other. WiBi is created in three steps: we first create a taxonomy for the Wikipedia pages by parsing textual definitions, extracting the hypernym(s) and disambiguating them according to the page inventory; next, we leverage the hypernyms in the page taxonomy, together with their links to the corresponding categories, so as to induce a taxonomy over Wikipedia categories while at the same time improving the page taxonomy in an iterative way; finally we employ structural heuristics to overcome inherent problems affecting categories. The output of our three-phase approach is a bitaxonomy of millions of pages and hundreds of thousands of categories for the English Wikipedia.

Acknowledgements

The author gratefully acknowledges the support of the ERC Starting Grant Multi-JEDI No. 259234.

References

Tiziano Flati, Daniele Vannella, Tommaso Pasini, and Roberto Navigli. 2014. Validating and Extending Semantic Knowledge Bases using Video Games with a Purpose. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pages 945–955, Baltimore, USA.

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.

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.

Daniele Vannella, David Jurgens, Daniele Scarfini, Domenico Toscani, and Roberto Navigli. 2014. Validating and Extending Semantic Knowledge Bases using Video Games with a Purpose. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pages 1294–1304, Baltimore, USA.