Three birds (in the LOD cloud) with one stone: BabelNet, Babelfy and the Wikipedia Bitaxonomy!

Tiziano Flati and Roberto Navigli

Dipartimento di Informatica
Sapienza Università di Roma

Abstract. In this paper we present the current status of linguistic resources published as linked data in the LLOD cloud in our research group, namely BabelNet, Babelfy and the Wikipedia Bitaxonomy. We describe the different resources in terms of their salient aspects, objectives and expected output and discuss the benefits that each of these resources potentially brings in the world of LLOD NLP-aware services. We also present public web-based services which allow to query, explore and export data into RDF format, thus truly enabling interoperability across the web.

1 Introduction

Recent years have witnessed a surge in the amount of semantic information published on the Web. Indeed, the Web of Data has been increasing steadily in both volume and variety, transforming the Web into a global database in which resources are linked across sites. It is becoming increasingly critical that existing lexical resources be published as Linked Open Data (LOD), so as to foster integration, interoperability and reuse on the Semantic Web [4]. Thus, lexical resources provided in RDF format can contribute to the creation of the so-called Linguistic Linked Open Data (LLOD, see Fig. 2), a vision fostered by the Open Linguistic Working Group (OWLG), in which part of the Linked Open Data cloud is made up of interlinked linguistic resources [1].

The multilinguality aspect is key to this vision, in that it enables Natural Language Processing tasks which are not only cross-lingual, but also independent both of the language of the user input and of the linked data exploited to perform the task. Both Semantic Web and Natural Language Processing communities face a new challenge, i.e., that of facilitating multilingual access to the Web of data.

The benefits of such a Web of Linguistic Data are diverse and lie on both Semantic Web and NLP sides. On the one hand, ontologies and linked data sets can be augmented with rich linguistic information, thereby enhancing Web-based information processing. On the other hand, NLP algorithms can take advantage of the availability of a vast, interoperable and federated set of linguistic resources, as well as benefit from a rich ecosystem of formalisms and technologies.

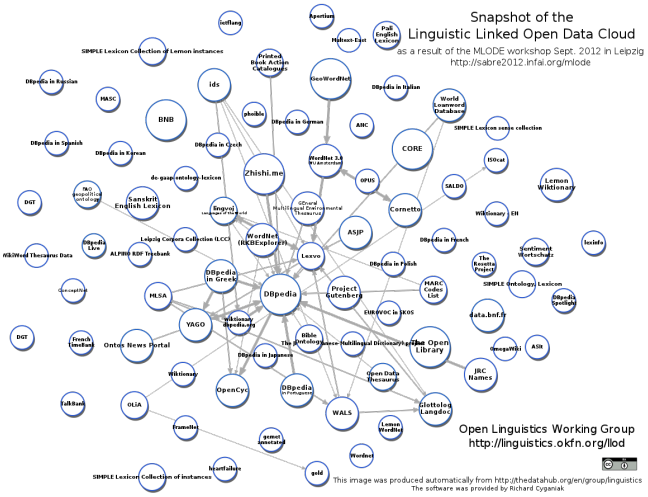


Fig. 1. The LLOD cloud.

This paper presents a contribution for the Multilingual Web of Data, with the publication of BabelNet 2.0, Babelfy and the Wikipedia Bitaxonomy as linked data. We describe the different resources in terms of their salient aspects, objectives and expected output and discuss the benefits that each of these resources potentially brings in the world of LLOD NLP-aware services.

2 The three resources in the cloud

We will now describe the major three resources oriented to the Linguistic Linked Open Data Cloud developed in our research group. These three resources, despite being different in nature as well as in their goals, they all have in common the linked data layer that allows the interlinking of information across the resources.

BabelNet BabelNet [7] is a very large multilingual encyclopedic dictionary and ontology whose version 2.0 covers 50 languages. Based on the integration of lexicographic and encyclopedic knowledge, BabelNet 2.0 offers a large network of concepts and named entities along with an extensive multilingual lexical coverage. The last version of BabelNet (as of writing, 2.5) is available at `babelnet.org` and a SPARQL endpoint is also accessible at `babelnet.org:8084/sparql/`. Lemon-BabelNet features more than 1 billion triples which describe 9.3 million concepts with encyclopedic and lexical information in 50 languages. The resource is interlinked with several other datasets including DBpedia as nucleus of the LOD cloud.

Babelfy The current language explosion on the Web requires the ability to automatically analyze and understand text written in any language. This task however is affected by the lexical ambiguity of language, an issue addressed

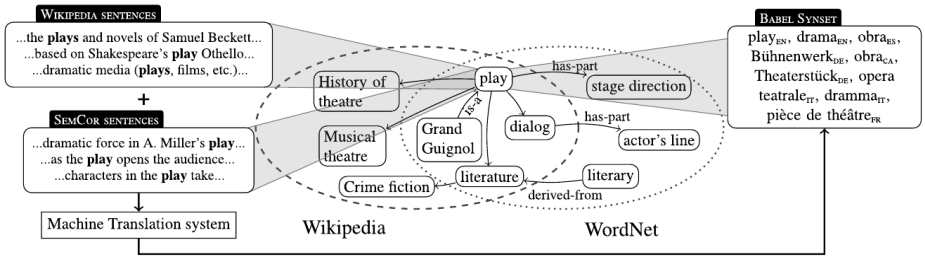


Fig. 2. BabelNet 2.5 overview.

by two key tasks: Multilingual Word Sense Disambiguation (WSD), aimed at assigning meanings to word occurrences within text, and Entity Linking (EL), a recent task focused on finding mentions of entities within text and linking them to a knowledge base.

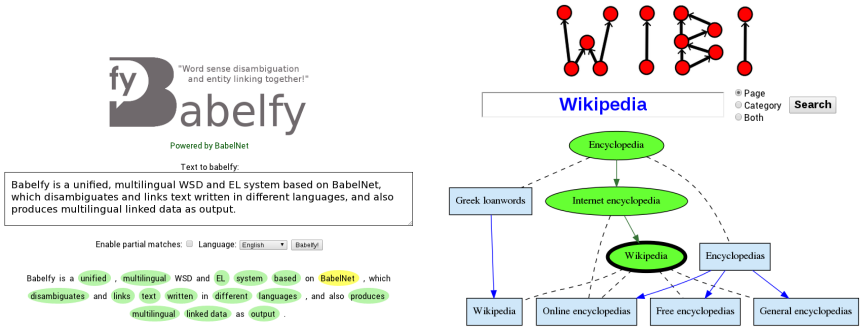
Babelfy [6] is a unified, multilingual WSD and EL system based on BabelNet, which disambiguates and links text written in different languages, and also produces multilingual linked data as output. Its peculiarity consists in the fact that, by combining a loose candidate identification and a novel densest graph heuristic, this system fares well both on long texts, such as those of the WSD tasks, and short sentences, such as the ones in EL tasks, thus indeed bringing together the best of the two worlds. Experiments conducted on six gold-standard datasets used in WSD and EL tasks show that Babelfy provides state-of-the-art results both in monolingual and multilingual setting. Babelfy also comes with RESTful APIs which programmatically let users retrieve disambiguated text with a few Java lines. An online live version of Babelfy is accessible at `babelfy.org`.

The Wikipedia Bitaxonomy The Wikipedia Bitaxonomy, also known as WiBi, is a project which aims at automatically extracting two taxonomies, one for Wikipedia pages and one for Wikipedia categories, aligned to each other, in a joint fashion with state-of-the-art results (see [3] for details).

Extensive comparison has been carried out on two datasets of 1,000 pages and categories each, against all the available knowledge resources, including MENTA, DBpedia, YAGO, WikiTaxonomy and WikiNet [5]. Results show that WiBi overcomes all competitors not only in terms of quality, with the highest precision and recall, but also in terms of coverage and specificity.

WiBi is also integrated into BabelNet and explorable through a web application at `wibitaxonomy.org`. Backed by the Apache Jena framework, the explorer integrates a single-click functionality that seamlessly converts the displayed data into RDF format, in line with recent work on linguistic linked open data and the Semantic Web (see [2]). The user can opt for Turtle, RDF/XML or N-Triple format.

Even though WiBi is (still) monolingual, recent research suggests that around 50% of Wikipedia in any other language is automatically covered for free, by simply projecting the English bitaxonomy through interlanguage links.



(a) Babelfy web application. (b) WiBi web application.

Fig. 3. The Wikipedia Bitaxonomy Explorer overview.

3 Conclusions

We described three resources that seamlessly integrate linked data facilities and thus foster interoperability within the LLOD cloud. Despite addressing different goals and offering different services, all of the three resources export data into RDF format and thus enable NLP-aware services to consume and re-elaborate data through the Semantic Web. If carefully published and interlinked, these resources could, indeed, potentially turn into a huge body of machine-readable knowledge and move on towards a full-fledged linguistic linked open cloud.

References

1. Chiarcos, C., Hellmann, S., Nordhoff, S.: Towards a Linguistic Linked Open Data Cloud: The Open Linguistics Working Group. TAL 52(3), 245–275 (2011)
2. Ehrmann, M., Cecconi, F., Vannella, D., Mccrae, J.P., Cimiano, P., Navigli, R.: Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0. In: Proc. of LREC 2014. European Language Resources Association (ELRA), Reykjavik, Iceland
3. Flati, T., Vannella, D., Pasini, T., Navigli, R.: Two Is Bigger (and Better) Than One: the Wikipedia Bitaxonomy Project. In: Proc. of ACL 2014. pp. 945–955. Baltimore, Maryland
4. Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., McCrae, J.: Challenges for the multilingual web of data. J. Web Sem. 11, 63–71 (2012)
5. Hovy, E.H., Navigli, R., Ponzetto, S.P.: Collaboratively built semi-structured content and Artificial Intelligence: The story so far. Artificial Intelligence 194, 2–27 (2013)
6. Moro, A., Raganato, A., Navigli, R.: Entity linking meets word sense disambiguation: a unified approach. TACL 2, 231–244 (2014)
7. Navigli, R., Ponzetto, S.P.: BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence 193, 217–250 (2012)