Three Birds (in the LLOD 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 and linguistic services in the LLOD cloud in our research group, namely BabelNet, Babelfy and the Wikipedia Bitaxonomy. We describe them in terms of their salient aspects and objectives and discuss the benefits that each of these potentially brings to the world of LLOD NLP-aware services. We also present public Web-based services which enable querying, exploring and exporting data into RDF format.

1 Introduction

Recent years have witnessed an upsurge 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 [5]. Thus, lexical resources provided in RDF format can contribute to the creation of the so-called Linguistic Linked Open Data (LLOD), 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 [2].

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 the Semantic Web and Natural Language Processing communities have to face the new challenge 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.

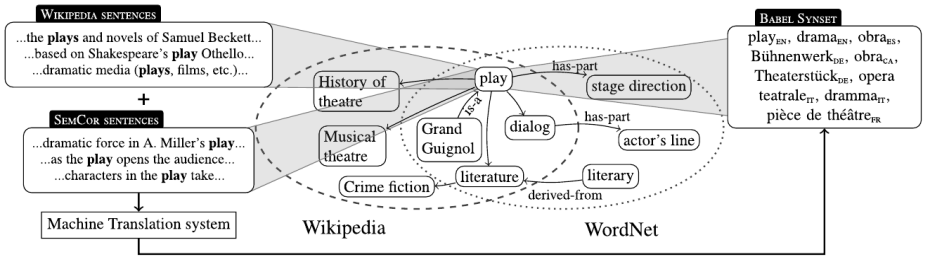


Fig. 1. BabelNet overview (picture from [10]).

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

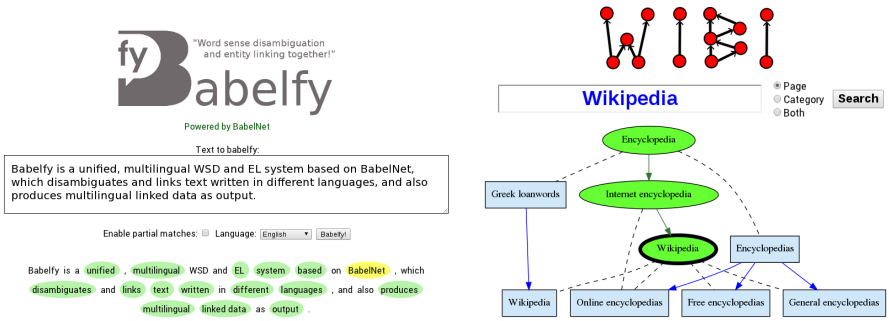
2 Three birds in the LLOD cloud

We now describe the three major tools and resources oriented to the Linguistic Linked Open Data Cloud developed in our research group. Despite being different in nature as well as in their goals (Babelfy is a service while BabelNet and the Wikipedia Bitaxonomy are linguistic resources), they all have in common the linked data layer that enables the interlinking of information across entities.

The three services, already useful on their own, are closely intertwined and beneficial to each other: in fact, while on the one hand the BabelNet semantic network lies at the core of Babelfy, on the other hand the Wikipedia Bitaxonomy is also integrated into BabelNet and acts as the taxonomical backbone of the resource.

BabelNet BabelNet [10] is a very large multilingual encyclopedic dictionary and ontology which covers 50 languages. Based on the automatic integration of lexicographic and encyclopedic knowledge extracted from multiple resources (WordNet, Wikipedia, Open Multilingual WordNet, OmegaWiki, Wiktionary and WikiData), BabelNet offers a large network of concepts and named entities along with an extensive multilingual lexical coverage (see Fig. 1). The last version of BabelNet is available at `babelnet.org` and a SPARQL endpoint is also accessible at `babelnet.org:8084/sparql/`. Based on the *lemon* model [7], a lexicon model for representing and sharing ontology lexica on the Semantic Web, the RDF-version of BabelNet (*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 and *lemon*-WordNet, thus laying the foundations for further linked data-based integration of ontology lexica.

Babelfy The current language explosion on the Web requires the ability to automatically analyze and understand text written in any language. This task



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

Fig. 2. Screenshots of Babelfy (a) and the Wikipedia Bitaxonomy Explorer (b).

however is affected by the lexical ambiguity of language, an issue addressed by two key tasks: Multilingual Word Sense Disambiguation (WSD) [1, 9], aimed at assigning meanings to word occurrences within text, and Entity Linking (EL) [11], a recent task focused on finding mentions of entities within text and linking them to a knowledge base.

Babelfy [8] 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 (see Fig. 2(a)). At its core are the combination of a loose candidate identification with a novel densest graph heuristic. Babelfy fares well both on long texts, such as those of the WSD tasks, and short sentences, such as the ones in EL tasks, thus 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 settings. Babelfy also comes with a RESTful API which programmatically enables users to retrieve disambiguated text with a few Java lines. An online 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 [4] for details).

Extensive comparison has been carried out on two datasets of 1,000 pages and categories respectively, against all the available knowledge resources, including MENTA, DBpedia, YAGO, WikiTaxonomy and WikiNet (see [6] for a comprehensive survey). 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` (see Fig. 2(b)). Backed by the Apache Jena framework, the explorer integrates a single-click functionality that seamlessly converts the displayed data into RDF format (either Turtle, RDF/XML or N-Triple), in line with recent work on LLOD and the Semantic Web (see [3]).

3 Conclusions

We described resources and services that seamlessly integrate linked data facilities and thus foster interoperability within the LLOD cloud, also across languages. Despite addressing different goals and offering different services, all of the three tools 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 tools could, indeed, potentially turn into a huge body of machine-readable knowledge and move on towards a full-fledged linguistic linked open data cloud.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Starting Grant MultiJEDI No. 259234.

The authors also acknowledge support from the LIDER project (No. 610782), a Coordination and Support Action funded by the European Commission under FP7.

References

1. Banea, C., Mihalcea, R.: Word Sense Disambiguation with Multilingual Features. In: Proc. of IWCS 2011. pp. 25–34. Association for Computational Linguistics, Stroudsburg, PA, USA
2. Chiarcos, C., Hellmann, S., Nordhoff, S.: Towards a Linguistic Linked Open Data Cloud: The Open Linguistics Working Group. TAL 52(3), 245–275 (2011)
3. 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. pp. 401–408. Reykjavik, Iceland
4. 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
5. 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)
6. 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)
7. McCrae, J., Montiel-Ponsoda, E., Cimiano, P.: Collaborative semantic editing of linked data lexica. In: Proc. of LREC 2012. pp. 2619–2625. Istanbul, Turkey
8. Moro, A., Raganato, A., Navigli, R.: Entity Linking meets Word Sense Disambiguation: a Unified Approach. TACL 2, 231–244 (2014)
9. Navigli, R., Ponzetto, S.P.: Joining Forces Pays Off: Multilingual Joint Word Sense Disambiguation. In: Proc. of EMNLP-CoNLL 2012. pp. 1399–1410. Jeju Island, Korea
10. 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)
11. Rao, D., McNamee, P., Dredze, M.: Entity Linking: Finding Extracted Entities in a Knowledge Base. In: Multi-source, Multi-lingual Information Extraction and Summarization, pp. 93–115 (2013)