(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 – 00161 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 we are going to overview and showcase in this paper.
We start by presenting BabelNet 2.5 [21,22], available at http://babelnet.org, a very large multilingual encyclopedic dictionary and semantic network, which currently 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.
Next, we present Babelfy [17], available at http://babelfy.org, a unified approach that leverages BabelNet to jointly perform word sense disambiguation 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.
Finally we describe two approaches to large-scale knowledge acquisition and validation: video games with a purpose [11,33], a novel, powerful paradigm for the large-scale acquisition and validation of knowledge and data, implemented as an online platform (http://knowledgeforge.org), and WiBi [7], available at http://wibitaxonomy.org, our approach to the construction of a Wikipedia bitaxonomy, that is, the largest and most accurate currently available taxonomy of Wikipedia pages and taxonomy of categories, aligned to each other.

1 Introduction

Integrating Knowledge. The creation of very large knowledge bases has been made possible by the availability of collaboratively-curated online resources such as Wikipedia and Wiktionary [14]. These resources are increasingly becoming enriched with new content in many languages and, although they are only partially structured, they provide a great deal of valuable knowledge which can be harvested and transformed into structured form [10,23]. Prominent examples include DBpedia [2], YAGO [9] and WikiNet [18]. However, these resources are mostly focused on the encyclopedic side of knowledge, overshadowing the lexicographic side which is covered in dictionaries. It is this weakness that we addressed in BabelNet [21,22], thanks to the automatic creation of

a large multilingual semantic network which integrates both lexicographic and encyclopedic knowledge from several different resources, including WordNet, Wikipedia and Wiktionary. BabelNet preserves the organizational structure of WordNet, i.e., it encodes concepts and named entities as sets of synonyms (synsets), but also takes it to the next level: the lexicalizations within synsets are available in multiple languages and the lexicographic information typical of WordNet is complemented with wide encyclopedic coverage, resulting in an intertwined network of concepts and named entities. We overview BabelNet in Section 2.

Disambiguating with Knowledge. Recent years have witnessed a surge in the amount of text published on the Web in a wide variety of languages. Being able to understand this text automatically at the semantic level is key, in that it enables Natural Language Processing tasks which are not only cross-lingual, but also independent of the language of the user input and the data exploited to perform the task. In order to enable automatic text understanding, Word Sense Disambiguation (WSD) [19,20] has to be performed. WSD is a historical task aimed at assigning meanings to single-word and multi-word occurrences within text, a task which is more alive than ever in the research community. In fact, the collaborative creation of large semi-structured resources has favoured the emergence of new tasks, such as Entity Linking (EL) [29], and opened up new possibilities for tasks such as Named Entity Disambiguation (NED) and Wikification. The aim of EL is to discover encyclopedic mentions of entities within a text and to link them to the most suitable entry in a reference knowledge base. Unfortunately, creating large amounts of training data for all possible meanings is not a feasible task, unless we resort to pseudowords [25], not to mention the multilingual aspect of such an effort. As a result, supervised systems are not suitable to this task.

In our laboratory, we recently proposed a novel wide-coverage graph-based algorithm, called Babelfy [17], for performing WSD and EL at the same time in arbitrary languages with state-of-the-art performance. We present Babelfy in Section 3.

Structuring Knowledge. Unlike the case with smaller manually-curated resources such as WordNet [6], in many large automatically-created resources the taxonomical information is either missing, mixed across resources, e.g., linking Wikipedia categories to WordNet synsets as in YAGO, or coarse-grained as is the case in DBpedia, whose hypernyms link to a small upper taxonomy.

Current automatic approaches in the literature have mostly focused on the extraction of taxonomies from the network of Wikipedia categories. WikiTaxonomy [28], the first approach of this kind, was based on the use of heuristics to determine whether is-a relations hold between a category and its subcategories. Subsequent approaches have also exploited heuristics, but have extended them to any kind of semantic relation expressed in the category names [18]. But while the above-mentioned attempts provide structure for categories, surprisingly little attention has been paid to the acquisition of a taxonomy for Wikipedia pages themselves. For instance, [30] provided a vector-based method which, however, was incapable of linking pages which do not have a WordNet counterpart. Higher coverage is provided by [15] thanks to the use of a set of effective heuristics, however, the approach also relies on WordNet and sense frequency information.

In our laboratory we tackled the task of taxonomizing Wikipedia in a way that is fully independent of other existing resources such as WordNet. The output of our research, WiBi [7], is a novel approach to the creation of a Wikipedia bitaxonomy, that is, a taxonomy of Wikipedia pages aligned to a taxonomy of categories. At the core of our approach lies the idea that the information at the page and category level are mutually beneficial for inducing a wide-coverage and fine-grained integrated taxonomy. We present WiBi in Section 4.

Playing with Knowledge. Not only do our systems inherently make mistakes in integrating lexical-semantic resources and in disambiguating text, but virtually all of Natural Language Processing depends on annotated examples. Possible options to address the annotation bottleneck are crowdsourcing [32] and games with a purpose [4,24]. However, while gamifying an annotation task has been shown to yield better quality and higher user engagement [12], most of the proposed games are inherently text-based, i.e., closer to a traditional annotation task than to video games people typically play.

In our laboratory, we introduced a novel paradigm for gamifying annotation tasks based on video games with a purpose, that is, games with graphical, dynamic features. Key to our approach are the development of a video game that is playable alone and the seamless integration of the annotation task into the game as a central component. The underlying idea is that, by focusing on the game play, players feel engaged and provide higher-quality annotations. We describe this novel paradigm in Section 5.

2 BabelNet

BabelNet (`http://babelnet.org`) is a large-scale encyclopedic dictionary and semantic network which encodes knowledge as a labeled directed graph $G = (V, E)$ where V is the set of *nodes* – i.e., *concepts* such as play and *named entities* such as Shakespeare – and $E \subseteq V \times R \times V$ is the set of *edges* connecting pairs of concepts (e.g., play *is-a* dramatic composition). Each edge is labeled with a *semantic relation* from R, e.g., {*is-a*, *part-of*, ..., ϵ}, where ϵ denotes an unspecified semantic relation. Importantly, each node $v \in V$ contains a set of lexicalizations of the concept for different languages, e.g., { play$_{\text{EN}}$, Theaterstück$_{\text{DE}}$, dramma$_{\text{IT}}$, obra$_{\text{ES}}$, ..., pièce de théâtre$_{\text{FR}}$ }. We call such multilingually lexicalized concepts *Babel synsets*. At its core, concepts and relations in BabelNet are harvested from the largest available semantic lexicon of English, WordNet, and a wide-coverage collaboratively-edited encyclopedia, Wikipedia. In order to construct the BabelNet graph, 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. Thus, our methodology consists of three main steps:

1. We **integrate WordNet and Wikipedia** by automatically creating a mapping between WordNet senses and Wikipages. This avoids duplicate concepts and allows their inventories to complement each other.
2. We **collect multilingual lexicalizations** of the newly-created concepts (i.e., Babel synsets) by using (a) the human-generated translations provided by Wikipedia (i.e., the *inter-language* links), as well as (b) a machine translation system to translate occurrences of the concepts within sense-tagged corpora.
3. We **create relations between Babel synsets** by harvesting all the relations in WordNet and in the wikipedias in the languages of interest.

Starting from this core, thanks to pairwise automatic mapping algorithms [26], the current version, i.e., BabelNet 2.5, integrates the following resources:

- WordNet [6], a popular computational lexicon of English (version 3.0),
- Wikipedia,[1] the largest collaborative multilingual Web encyclopedia (October 2012 dumps),
- Open Multilingual WordNet [3], a collection of wordnets available in different languages (August 2013 dump),
- OmegaWiki,[2] a large collaborative multilingual dictionary (01/09/2013 dump).
- Wiktionary,[3] a large collaborative dictionary (11/03/2014 dump).
- Wikidata,[4] a collaborative project for creating a large knowledge base (20/04/2014 dump).

BabelNet is available online not only as (i) a public user interface for human consumption, but also as LLOD (Linguistic Linked Open Data) [5] using the lemon-RDF model [13] via: (ii) a public SPARQL endpoint set up using the Virtuoso Universal Server; (iii) dereferenceable URIs available via the Pubby Linked Data SPARQL frontend.

3 Babelfy

Babelfy (`http://babelfy.org`) is a unified graph-based approach to EL and WSD based on a loose identification of candidate meanings coupled with a densest subgraph heuristic which selects high-coherence semantic interpretations [17]. Babelfy works in three steps:

1. 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.
2. 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.

[1] `http://wikipedia.org`
[2] `http://omegawiki.org`
[3] `http://wiktionary.org`
[4] `http://wikidata.org`

3. 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 tasks on 6 different gold-standard datasets, including a multilingual setting. An online API is available for querying Babelfy in any of the 50 languages currently available in BabelNet [16].

4 WiBi: The Wikipedia Bitaxonomy

In order to provide a taxonomical backbone for the entire BabelNet (instead of limiting it to the WordNet subgraph only), we proposed a method for producing hypernym (i.e., is-a) relations for most of the Wikipedia pages. At the core of our approach lies the idea that the information at the page and category level are mutually beneficial for inducing a wide-coverage and fine-grained integrated taxonomy. We induce a Wikipedia Bitaxonomy (WiBi) [7], i.e., a taxonomy of pages and categories, in 3 phases:

1. **Creation of the initial page taxonomy:** as a first step, we create a taxonomy for the Wikipedia pages by parsing textual definitions, extracting the hypernym(s) and disambiguating them according to the page inventory.
2. **Creation of the bitaxonomy:** 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 in an iterative way. At each iteration, the links in the page taxonomy are used to identify category hypernyms and, conversely, the new category hypernyms are used to identify more page hypernyms.
3. **Refinement of the category taxonomy:** finally we employ structural heuristics to overcome inherent problems affecting categories.

The output of our three-phase approach, available at `http://wibitaxonomy.org`, is a bitaxonomy of millions of pages and hundreds of thousands of categories for the English Wikipedia. Our experiments show that WiBi provides a richer and more accurate structure than those produced in the literature, with nearly full coverage of pages and categories. While BabelNet currently covers Wikipedia pages only, in the near future we also plan to integrate categories together with their is-a relations from WiBi.

5 Video Games with a Purpose

Automatic systems make unavoidable mistakes, e.g. when integrating different lexical-semantic resources or disambiguating text. Moreover, the knowledge acquisition bottleneck hampers the creation and enrichment of datasets and knowledge bases. To address these two questions, that is, the validation and annotation of linguistic items, we proposed using **video games with a purpose** [11,33] as a novel paradigm inspired by the

idea of games with a purpose [1]. Here, the annotation tasks are transformed into elements of a video game where players accomplish their jobs by virtue of playing the game, rather than by performing a more traditional annotation task. While prior efforts in NLP have incorporated games for performing annotation and validation [31,8,27], these games have largely been text-based, adding game-like features such as high-scores on top of an existing annotation task. In contrast, we introduce video games with graphical 2D gameplay that is similar to what game players are familiar with. The fun nature of the games provides an intrinsic motivation for players to keep playing, which can increase the quality of their work and lower the annotation cost.

Our work provides the following contributions:

- We demonstrate effective video game-based methods for both validating and extending semantic networks [33], and for disambiguating text [11], using video games that operate on complementary sources of information: semantic relations and sense-image mappings.
- In contrast to previous work, the annotation quality can be determined in a fully automatic way thanks to the use of automatically-extracted negative items and appropriate game play mechanisms [33].
- We demonstrate that converting games with a purpose into more traditional video games creates an increased player incentive such that players annotate for free, thereby significantly lowering annotation costs below that of crowdsourcing.
- We show that games produce better quality annotations than crowdsourcing.

Our video game with a purpose paradigm is implemented as an online platform (`http://knowledgeforge.org`) where games can be supplied by developers and played by users who contribute annotations.

6 Conclusion

It is an exciting time for Natural Language Processing! Multilinguality is now addressed both in terms of automatic lexical-semantic resource building and integration and for high-performance disambiguation and entity linking. Moreover, novel paradigms, such as video games with a purpose, provide effective ways to contribute large amounts of annotations and validations, thereby bringing down the cost of such tasks and increasingly reducing the inevitable error rate intrinsic to automatic techniques.

Acknowledgements


The author gratefully acknowledges the support of the ERC Starting Grant MultiJEDI No. 259234.


References

1. von Ahn, L.: Games with a purpose. IEEE Computer 6(39), 92–94 (2006)
2. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia - a crystallization point for the Web of Data. Web Semantics 7(3), 154–165 (2009)
3. Bond, F., Foster, R.: Linking and extending an Open Multilingual Wordnet. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1352–1362. Association for Computational Linguistics, Sofia, Bulgaria (August 2013)
4. Chamberlain, J., Fort, K., Kruschwitz, U., Lafourcade, M., Poesio, M.: Using games to create language resources: Successes and limitations of the approach. In: Gurevych, I., Kim, J. (eds.) The People's Web Meets NLP, pp. 3–44. Theory and Applications of Natural Language Processing, Springer (2013)
5. 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: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014. pp. 401–408 (2014)
6. Fellbaum, C. (ed.): WordNet: An Electronic Database. MIT Press, Cambridge, MA (1998)
7. Flati, T., Vannella, D., Pasini, T., Navigli, R.: Two Is Bigger (and Better) Than One: the Wikipedia Bitaxonomy Project. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014). pp. 945–955. Baltimore, USA (2014)
8. Herdağdelen, A., Baroni, M.: Bootstrapping a game with a purpose for common sense collection. ACM Transactions on Intelligent Systems and Technology 3(4), 1–24 (2012)
9. Hoffart, J., Suchanek, F.M., Berberich, K., Weikum, G.: YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Artificial Intelligence 194, 28–61 (2013)
10. 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)
11. Jurgens, D., Navigli, R.: It's all fun and games until someone annotates: Video games with a purpose for linguistic annotation. Transactions of the Association for Computational Linguistics (TACL) to appear (2014)
12. Lee, T.Y., Dugan, C., Geyer, W., Ratchford, T., Rasmussen, J., Shami, N.S., Lupushor, S.: Experiments on motivational feedback for crowdsourced workers. In: Seventh International AAAI Conference on Weblogs and Social Media. pp. 341–350 (2013)
13. McCrae, J., de Cea, G.A., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., et al.: Interchanging lexical resources on the semantic web. Language Resources and Evaluation 46(4), 701–719 (2012)
14. Medelyan, O., Milne, D., Legg, C., Witten, I.H.: Mining meaning from Wikipedia. International Journal of Human-Computer Studies 67(9), 716–754 (2009)
15. de Melo, G., Weikum, G.: MENTA: Inducing Multilingual Taxonomies from Wikipedia. In: Proceedings of Conference on Information and Knowledge Management (CIKM '10). pp. 1099–1108. New York, NY, USA (2010)
16. Moro, A., Cecconi, F., Navigli, R.: Multilingual Word Sense Disambiguation and Entity Linking for Everybody. In: Proc. of the International Semantic Web Conference (P&D) (2014)
17. Moro, A., Raganato, A., Navigli, R.: Entity Linking meets Word Sense Disambiguation: a Unified Approach. Transactions of the Association for Computational Linguistics (TACL) 2, 231–244 (2014)
18. Nastase, V., Strube, M.: Transforming Wikipedia into a large scale multilingual concept network. Artificial Intelligence 194, 62–85 (2013)

19. Navigli, R.: Word Sense Disambiguation: A survey. ACM Computing Surveys 41(2), 1–69 (2009)
20. Navigli, R.: A quick tour of Word Sense Disambiguation, induction and related approaches. In: Bieliková, M., Friedrich, G., Gottlob, G., Katzenbeisser, S., Turán, G. (eds.) SOFSEM 2012: Theory and Practice of Computer Science, Lecture Notes in Computer Science, vol. 7147, pp. 115–129. Heidelberg: Springer (2012)
21. Navigli, R., Ponzetto, S.P.: BabelNet: Building a very large multilingual semantic network. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010. pp. 216–225 (2010)
22. 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)
23. Navigli, R., Velardi, P.: From glossaries to ontologies: Extracting semantic structure from textual definitions. In: Buitelaar, P., Cimiano, P. (eds.) Ontology Learning and Population: Bridging the Gap between Text and Knowledge. pp. 71–87. IOS Press, Amsterdam, The Netherlands (2008)
24. Pe-Than, E.P.P., Goh, D.L., Lee, C.S.: A survey and typology of human computation games. In: Information Technology: New Generations (ITNG), 2012 Ninth International Conference on. pp. 720–725. IEEE (2012)
25. Pilehvar, M.T., Navigli, R.: A Large-scale Pseudoword-based Evaluation Framework for State-of-the-Art Word Sense Disambiguation. Computational Linguistics 40(4) (2014)
26. Pilehvar, M.T., Navigli, R.: A robust approach to aligning heterogeneous lexical resources. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014). pp. 468–478. Baltimore, Maryland (2014)
27. Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., Ducceschi, L.: Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. ACM Transactions on Interactive Intelligent Systems 3(1), 3:1–3:44 (Apr 2013)
28. Ponzetto, S.P., Strube, M.: Deriving a large scale taxonomy from Wikipedia. In: Proceedings of the 22nd Conference on the Advancement of Artificial Intelligence (AAAI '07), Vancouver, B.C., Canada, 22–26 July 2007. pp. 1440–1445 (2007)
29. Rao, D., McNamee, P., Dredze, M.: Entity Linking: Finding Extracted Entities in a Knowledge Base. In: Multi-source, Multilingual Information Extraction and Summarization, pp. 93–115. Theory and Applications of Natural Language Processing, Springer Berlin Heidelberg (2013)
30. Ruiz-Casado, M., Alfonseca, E., Castells, P.: Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets. In: Advances in Web Intelligence, Lecture Notes in Computer Science, vol. 3528, pp. 380–386. Springer Verlag (2005)
31. Siorpaes, K., Hepp, M.: Ontogame: Weaving the semantic web by online games. In: Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M. (eds.) The Semantic Web: Research and Applications. Lecture Notes in Computer Science, vol. 5021, pp. 751–766. Springer Berlin Heidelberg (2008)
32. Snow, R., O'Connor, B., Jurafsky, D., Ng, A.: Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. pp. 254–263 (2008)
33. Vannella, D., Jurgens, D., Scarfini, D., Toscani, D., Navigli, R.: 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). pp. 1294–1304. Baltimore, USA (2014)