BabelNetXplorer: A Platform for Multilingual Lexical Knowledge Base Access and Exploration

Roberto Navigli
Dipartimento di Informatica
Sapienza Università di Roma
Rome, Italy
navigli@di.uniroma1.it

Simone Paolo Ponzetto
Dipartimento di Informatica
Sapienza Università di Roma
Rome, Italy
ponzetto@di.uniroma1.it

ABSTRACT

Knowledge on word meanings and their relations across languages is vital for enabling semantic information technologies: in fact, the ever increasingly multilingual nature of the Web now calls for the development of methods that are both robust and widely applicable for processing textual information in a multitude of languages. In our research, we approach this ambitious task by means of BabelNet, a wide-coverage multilingual lexical knowledge base. In this paper we present an Application Programming Interface and a Graphical User Interface which, respectively, allow programmatic access and visual exploration of BabelNet. Our contribution is to provide the research community with easy-to-use tools for performing multilingual lexical semantic analysis, thereby fostering further research in this direction.

Categories and Subject Descriptors

I.2.4 [**Artificial Intelligence**]: Semantic Networks; H.3.4 [**Information Storage and Retrieval**]: Systems and Software; H.4.3 [**Information Systems Applications**]: Information Browsers

General Terms

Algorithms

Keywords

Semantic networks, visualization, multilinguality.

1. INTRODUCTION

The vast amount of textual content now on the Web opens up new challenges for Natural Language Processing (NLP), especially in terms of developing wide-coverage, domain-independent, multilingual applications. In fact, these huge repositories of text contain a great wealth of information, which can be harvested automatically to address the so-called knowledge acquisition bottleneck. Recently Web resources (including online collaborative efforts such as Wikipedia[1]) have been leveraged for the automatic acquisition of wide-coverage multilingual lexical knowledge resources in a weakly or fully unsupervised fashion [2, 6, 10]. These resources, in their turn, make it possible to generate and leverage semantically rich representations based on world and linguistic knowledge that are not only beneficial for complex NLP tasks such as Word Sense Disambiguation (WSD) [8, 12], but also enable next-generation systems embedding machine-readable knowledge within end-user applications, such as Machine Translation and Information Retrieval.

Our vision of knowledge-rich multilingual NLP requires two fundamental ingredients: first, a wide-coverage multilingual lexical knowledge base; second, a set of tools to query, retrieve and visualize information from this knowledge base in an effective manner. However, to date, there are no integrated resources and tools which are freely available for the research community on a multilingual scale. Previous endeavors are either not freely available (EuroWordNet [16]), or are only accessible via a Web interface (cf. the Multilingual Research Repository [1] and MENTA [2]), or provide only 'raw' data with no library for programmatic access (e.g. WikiNet [6]). However, the availability of easy-to-use libraries for efficient information access is known to foster top-level research – consider, for example, the widespread use of semantic similarity measures in NLP, due in no small measure to the availability of `WordNet::Similarity` [11]. Similarly, the availability of user interfaces such as WordNet's[2] or Visual Thesaurus[3] beneficially enables fast information access and browsing. With this paper we aim to fill this gap in availability of multilingual tools and resources, introducing BabelNetXplorer, a multi-tiered contribution consisting of (a) the full public release of BabelNet [10], a knowledge repository with concept lexicalizations in 6 languages (Catalan, English, French, German, Italian and Spanish); (b) an Application Programming Interface (API) to efficiently access both lexicographic (i.e., word senses) and conceptual (i.e., concepts and semantic relations) information found in BabelNet; (c) a graphical interface that allows the user to visually browse BabelNet's content.

2. BABELNET

BabelNet [10] follows the structure of a traditional lexical knowledge base and accordingly consists of a labeled directed graph where nodes represent concepts and named entities, and edges express semantic relations between them. Concepts and relations are harvested from the largest available semantic lexicon of English, i.e., WordNet [3], and a

[1] `http://www.wikipedia.org`

[2] `http://wordnetweb.princeton.edu/perl/webwn`

[3] `http://www.visualthesaurus.com`

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

```
bn:00008364n WIKIWN 08420278n 85 WN:EN:bank:1 WIKI:EN:Bank:2 WIKI:DE:Bank:2 WIKIRED:DE:Finanzinstitut:3
                                 WN:EN:banking_company:4 WIKI:IT:Banca:2 WNTR:ES:banco:1
                                 WNTR:FR:société_bancaire:4 WIKIRED:ES:Banca_telefonica:5 ...
                              228 r bn:02945246n FROM_IT|r bn:02854884n ~ bn:00044511n @ bn:00034537n ...
```

Figure 1: The Babel synset for bank^2_n, i.e. its 'financial' sense (excerpt, formatted for ease of readability).

wide-coverage collaboratively-edited encyclopedia, i.e., Wikipedia, thus making BabelNet a multilingual 'encyclopedic dictionary' which combines lexicographic information with wide-coverage encyclopedic knowledge. BabelNet's concept inventory consists of all WordNet's word senses (e.g., bank^2_n)[4] and Wikipedia's encyclopedic entries (i.e., its pages, such as for instance BANK OF AMERICA[5]), while its set of available relations comprises both semantic pointers between WordNet synsets (bank^2_n *is-a* financial institution^1_n) and semantically unspecified relations from Wikipedia's hyperlinked text (BANK OF AMERICA *is-related-to* RECESSION). In addition to this conceptual backbone, BabelNet provides a multilingual lexical dimension. Each of its nodes, called *Babel synsets*, contains a set of lexicalizations of the concept in different languages, e.g., { bank$_{\text{EN}}$, Bank$_{\text{DE}}$, banca$_{\text{IT}}$, ..., banco$_{\text{ES}}$ }. Multilingual lexicalizations for all concepts are collected from Wikipedia's inter-language links (e.g., the English Wikipedia page BANK links to the Italian BANCA), as well as by filling the translation gaps (i.e., missing translations) by means of a statistical machine translation system applied to sense-tagged data from SemCor [5] and Wikipedia itself – for instance, most occurrences of bank^1_n in SemCor are translated into German as Ufer.

3. BABELNETXPLORER

BabelNet dump. Similarly to WordNet, BabelNet can be stored in a plain text file. This file consists of a list of records, each one identifying a single Babel synset and represented in the following format:

```
id region offset sense-no sense+ relation-no relation+
```

An excerpt of the entry for the Babel synset containing bank^2_n is shown in Figure 1. The record contains (a) the synset's `id`; (b) the `region` of BabelNet where it lies (e.g., `WIKIWN` means at the intersection of WordNet and Wikipedia); (c) the corresponding (possibly empty) WordNet 3.0 synset `offset`; (d) the number of *senses* (i.e., `sense-no`) in all languages and their full listing (i.e., `sense+`); (e) the number of semantic pointers to other Babel synsets (i.e., `relation-no`) and their full listing (i.e., `relation+`). Senses encode information about their source – i.e., whether they come from WordNet (`WN`), Wikipedia pages (`WIKI`) or their redirections (`WIKIRED`), or are automatic translations (`WNTR` / `WIKITR`) – and about their language and lemma. In addition, senses are indexed to model intra-synset relations of translation across languages, i.e., co-indexed senses are translations from English to another language (for instance, bank^2_n translates to banca in Italian). Finally, semantic relations are encoded using WordNet's pointers, and an additional symbol for Wikipedia relations (`r`), which can also specify the source of the relation (e.g., `FROM_IT` means that the relation was harvested from the Italian Wikipedia). In the example in Figure 1, the Babel synset inherits WordNet hypernyms (`@`) and hyponyms (`~`) relations to financial institution^1_n (offset `bn:00034537n`) and Home Loan Bank^1_n (`bn:00044511n`), respectively, as well as Wikipedia relations to the synsets of FINANCIAL INSTRUMENT (`bn:02945246n`) and ETHICAL BANKING (`bn:02854884n`, from Italian).

BabelNet API. Information encoded in the text dump of BabelNet can be effectively accessed and automatically embedded within applications by means of a programmatic access. To this end, we developed a Java API, based on Apache Lucene[6] as backend, which indexes the textual dump and includes a variety of methods to access the three main levels of information encoded in BabelNet, namely: (a) lexicographic (information about word senses); (b) conceptual (the semantic network made up of its concepts); (c) and multilingual level (information about word translations). Figure 2 shows a usage example of our API, together with its output. The snippet starts by retrieving all the Babel synsets for the English word bank (line 3). Next, we access different kinds of information for each synset: first, in lines 5–7 we print its id, source (WordNet, Wikipedia, or both), corresponding WordNet offset (possibly empty), and 'main lemma' – namely, a compact string representation of the Babel synset consisting of its corresponding WordNet synset in stringified form, or the first non-redirection Wikipedia page found in it. Then, we access and print the German word senses of the Babel synsets (lines 8–10), and finally the synsets they are related to (lines 11–19). Note that thanks to carefully designed Java classes, we are able to accomplish all of this in about 20 lines of code.

Graphical user interface. We ship the API with a graphical user interface (GUI) that allows the user to visually browse the knowledge repository. Snapshots of the GUI are shown in Figure 3. The GUI has two main visualization modalities. The first (Figure 3(a)) allows the user to input an arbitrary list of words (top right pane) and build a semantic graph (i.e., a subgraph of BabelNet containing their senses and paths that connect pairs of senses [9]). We use a tree layout for visualization, as this allows for intuitive navigation. This visualization strategy, aimed at building semantic networks from an arbitrary set of words, is useful for exploring portions of BabelNet when the words of interest are known *a priori* (for instance, when we need to disambiguate words within a sentence). However, a user might wish instead to start from a specific word, select one of its senses and then its semantically related concepts. Accordingly, in order to allow a radial exploration of BabelNet, namely depth-first-search-like from a single sense, we designed a second visualization modality (Figure 3(b)). In

[4]We denote WordNet senses with w^i_p the i-th sense of a word w with part of speech p.

[5]We refer to Wikipedia pages and senses using SMALL CAPS.

[6]`http://lucene.apache.org`

```
BabelNet bn = BabelNet.getInstance();
System.out.println("SYNSETS WITH English word: \"bank\"");
List<BabelSynset> synsets = bn.getSynsets(Language.EN, "bank");
for (BabelSynset synset : synsets) {
  System.out.print("  =>(" + synset.getId() + ") SOURCE: " + synset.getSource() +
                   "; WN SYNSET: " + synset.getWordNetOffsets() + ";\n" +
                   "  MAIN LEMMA: " + synset.getMainLemma() + ";\n  SENSES (German): { ");
  for (BabelSense sense : synset.getSenses(Language.DE))
    System.out.print(sense.toString()+" ");
  System.out.println("}\n  -----");
  Map<IPointer, List<BabelSynset>> relatedSynsets = synset.getRelatedMap();
  for (IPointer relationType : relatedSynsets.keySet()) {
    List<BabelSynset> relationSynsets = relatedSynsets.get(relationType);
    for (BabelSynset relationSynset : relationSynsets) {
      System.out.println("    EDGE " + relationType.getSymbol() +
                         " " + relationSynset.getId() +
                         " " + relationSynset.toString(Language.EN));
    }
  }
  System.out.println("  -----");
}
```

```
SYNSETS WITH English word: "bank"
  ...
  =>(bn:00008364n) SOURCE: WIKIWN; WN SYNSET: [08420278n];
  MAIN LEMMA: depository_financial_institution#n#1|bank#n#2|banking_concern#n#1|banking_company#n#1;
  SENSES (German): { WIKI:DE:Bank WIKIRED:DE:Finanzinstitut WIKIRED:DE:Geschaeftsbanken ... WIKIRED:DE:Bankhaus }
    ------
    EDGE ~ bn:00020991n { Commercial_bank, Corporate_banking, ..., full_service_bank#n#1 }
    EDGE r bn:02945246n { Financial_instrument, Instrument_(finance), Liquid_financial_instrument }
  ...
  =>(bn:00008370n) SOURCE: WIKIWN; WN SYNSET: [04139859n];
  MAIN LEMMA: savings_bank#n#2|coin_bank#n#1|money_box#n#1|bank#n#8;
  SENSES (German): { WIKI:DE:Spardose WIKIRED:DE:Sparbüchse ... WIKIRED:DE:Sparschwein }
    ------
    EDGE r bn:00020497n { Coin_(money), Copper_coins, coin#n#1 }
```

Figure 2: Sample BabelNet API usage (with output).

this modality, starting from the node denoting an initial Babel synset, the user can recursively expand it by showing its outgoing edges and semantic relations to other concepts. The user can switch back and forth between the two views at any moment and visualize the selected relations in the context of the semantic graph or the radial network.

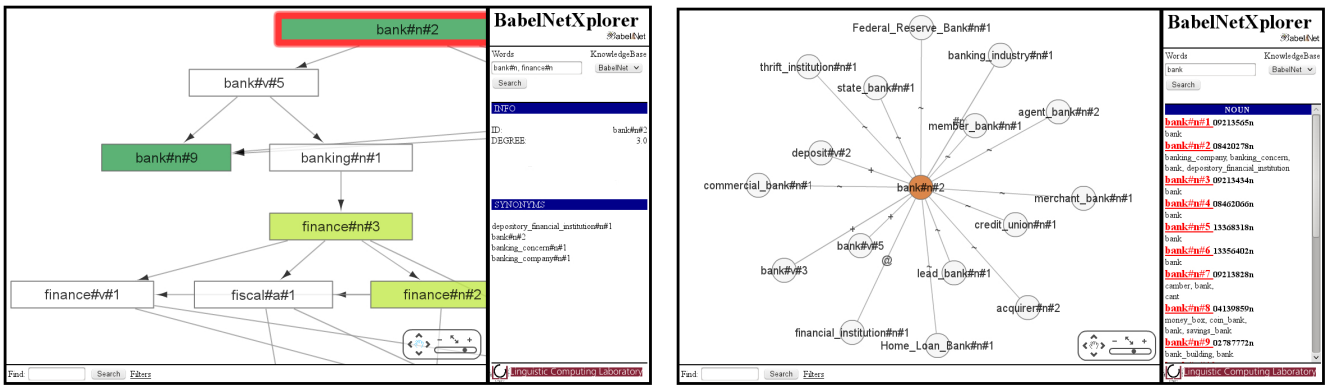
GUI: implementation details. To ensure robustness and scalability we make use of Cytoscape Web[7], a state-of-the-art network visualization software [13], which we pair with our own Java library to query paths and create semantic graphs with BabelNet. The latter works by pre-computing paths connecting any pair of Babel synsets, which are collected by iterating through each synset in turn, and performing a depth-first search up to a maximum depth – which we set to 3, on the basis of experimental evidence from a variety of knowledge base linking and lexical disambiguation tasks [9, 12]. Next, these paths are stored and indexed within a Lucene index, which ensures efficient lookups for querying those paths starting and ending in a specific synset. Given a set of words as input, a Java semantic graph factory class searches for their meanings within BabelNet, looks for their connecting paths, and merges such paths within a single graph. Optionally, the paths in the graph can be filtered – e.g., it is possible to remove loops, weighted edges below a certain threshold, etc. – and nodes in the final graph can be scored by means of a variety of methods – such as, for instance, their outdegree in the context of the semantic graph (cf. [9]). Note that the API and GUI's functionality are used in this work together with BabelNet: however, both can also be used in conjunction with other existing lexical knowledge resources, provided they can be wrapped around Java classes which implement interface methods for querying senses, concepts, and their semantic relations.

[7] http://cytoscapeweb.cytoscape.org

4. DEMONSTRATION

In the demonstration we show the two main contributions of this paper: (1) an API to programmatically access BabelNet; (2) the graphical user interface built on top of it.

We first introduce BabelNet and the problem of polysemy in different languages: for this purpose we start with words in arbitrary languages and show the relations between their different senses and translations across languages. The main objective of the session is to illustrate to the audience how to query BabelNet and navigate across its concepts and word senses in different languages. To this end, we make use of simple programs such as the one shown in Figure 2. While letting the users interactively query BabelNet's content, we guide them at the same time through the actual programs, in order to provide a walk-through of the API. In parallel, we show how the textual output of the programs is visually represented in the GUI. Using the visualization component, we show the audience how the previous text-based navigation can be more easily performed in a graphical environment. Moreover, thanks to the API's interfaces which can be applied to different lexical knowledge bases, we use the GUI to show the richness of BabelNet, which is compared side-by-side with WordNet.



(a) semantic graph / connectivity measures (b) thesaurus lookup / radial network exploration

Figure 3: BabelNetXplorer Graphical User Interface.

5. RELATED WORK

Recent work in knowledge base browsing and visualization has concentrated on building semantic graphs to perform Word Sense Disambiguation [7], summarization techniques for extracting semantic graphs expressing the most salient relations of an entity with its related concepts [15], as well as improving search by means of geographic and temporal information [4] – which is complementary to similar efforts in browsing document collections [14]. Our work complements these parallel contributions by means of an integrated platform (including both API and graphical components), which allows the user to query and search programmatically a very large multilingual lexical knowledge base, and to browse it visually. BabelNetXplorer builds upon BabelNet, a multilingual 'encyclopedic dictionary' bringing together the lexicographic and encyclopedic knowledge from WordNet and Wikipedia. Other recent efforts on creating multilingual knowledge bases from Wikipedia include WikiNet [6] and MENTA [2]: both these resources offer structured information complementary to BabelNet – i.e., large amounts of facts about entities (MENTA), and explicit semantic relations harvested from Wikipedia categories (WikiNet) – and will be integrated in the future into our API and GUI.

Acknowledgments

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

Thanks also go to Google for access to the University Research Program for Google Translate. BabelNet can be downloaded at `http://lcl.uniroma1.it/babelnet`.

6. REFERENCES

[1] J. Atserias, L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini, and P. Vossen. The MEANING multilingual central repository. In *Proc. of GWC-04*, pages 22–31, 2004.

[2] G. de Melo and G. Weikum. MENTA: Inducing multilingual taxonomies from Wikipedia. In *Proc. of CIKM-10*, pages 1099–1108, 2010.

[3] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Mass., 1998.

[4] J. Hoffart, F. M. Suchanek, K. Berberich, E. Lewis-Kelham, G. de Melo, and G. Weikum. YAGO2: Exploring and querying world knowledge in time, space, context, and many languages. In *Proc. of WWW-11*, pages 229–232, 2011.

[5] G. A. Miller, C. Leacock, R. Tengi, and R. Bunker. A semantic concordance. In *Proceedings of the 3rd DARPA Workshop on HLT*, pages 303–308, 1993.

[6] V. Nastase, M. Strube, B. Börschinger, C. Zirn, and A. Elghafari. WikiNet: A very large scale multi-lingual concept network. In *Proc. of LREC '10*, 2010.

[7] R. Navigli. Online word sense disambiguation with structural semantic interconnections. In *Proc. of EACL-06*, pages 107–110, 2006.

[8] R. Navigli. Word Sense Disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69, 2009.

[9] R. Navigli and M. Lapata. An experimental study on graph connectivity for unsupervised Word Sense Disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4):678–692, 2010.

[10] R. Navigli and S. P. Ponzetto. BabelNet: Building a very large multilingual semantic network. In *Proc. of ACL-10*, pages 216–225, 2010.

[11] T. Pedersen, S. Patwardhan, and J. Michelizzi. WordNet::Similarity – Measuring the relatedness of concepts. In *Comp. Vol. to Proc. of HLT-NAACL-04*, pages 267–270, 2004.

[12] S. P. Ponzetto and R. Navigli. Knowledge-rich Word Sense Disambiguation rivaling supervised systems. In *Proc. of ACL-10*, pages 1522–1531, 2010.

[13] M. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, and T. Ideker. Cytoscape 2.8: New features for data integration and network visualization. *Bioinformatics*, 27(3):431–432, 2011.

[14] J. Strötgen and M. Gertz. Timetrails: A system for exploring spatio-temporal information in documents. In *Proc. of VLDB 2010*, pages 1569–1572, 2010.

[15] T. Tylenda, M. Sozio, and G. Weikum. Einstein: physicist or vegetarian? Summarizing semantic type graphs for knowledge discovery. In *Proc. of WWW-11*, pages 273–276, 2011.

[16] P. Vossen, editor. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer, Dordrecht, The Netherlands, 1998.