BabelNet goes to the (Multilingual) Semantic Web

Roberto Navigli

Sapienza University of Rome, Via Salaria, 113 – 00198 Roma Italy,
navigli@di.uniroma1.it

Abstract. BabelNet is a very large, wide-coverage multilingual ontology. This resource is created by linking the largest multilingual Web encyclopedia – i.e., Wikipedia – to the most popular computational lexicon – i.e., WordNet. The integration is performed via an automatic mapping and by filling in lexical gaps in resource-poor languages with the aid of Machine Translation. The result is an "encyclopedic dictionary" that provides babel synsets, i.e., concepts and named entities lexicalized in many languages and connected with large amounts of semantic relations. BabelNet is available online at http://www.babelnet.org.
In this paper we present a first attempt at encoding BabelNet for the multilingual Semantic Web.

Keywords: lexicalized ontologies, semantic networks, multilinguality, lexical semantics

1 Introduction

In the information society, lexical knowledge is a key skill for understanding and decoding an ever-changing world. Indeed, lexical knowledge is an essential component not only for human understanding of text, but it is also indispensable for the creation of the multilingual Semantic Web. Unfortunately, however, building such lexical knowledge resources manually is an onerous task requiring dozens of years – and what is more it has to be repeated from scratch for each new language. On top of this, it is becoming increasingly critical that existing resources be published as Linked Open Data (LOD), so as to foster integration, interoperability and reuse on the Semantic Web [3].

Thus, lexical resources provided in RDF format [4] can contribute to the creation of the so-called Linguistic Linked Open Data (LLOD, see Figure 1), a vision fostered by the Open Linguistic Working Group (OWLG)[1] 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 of the language of the user input and the linked data exploited to perform the task.

[1] http://linguistics.okfn.org

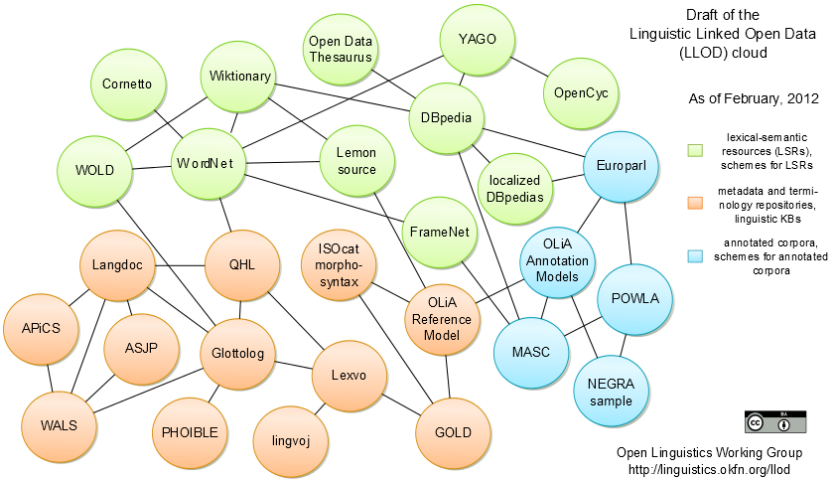

Fig. 1. Open Linguistics Working Group (2012), The Linguistic Linked Open Data cloud diagram (draft), version of February 2012, `http://linguistics.okfn.org/llod`.

This paper provides a contribution to the LLOD vision by presenting a first encoding of BabelNet in RDF. BabelNet (`http://www.babelnet.org`) is a very large multilingual semantic network obtained as a result of a novel integration and enrichment methodology. This resource is created by linking the largest multilingual Web encyclopedia – i.e., Wikipedia – to the most popular computational lexicon – i.e., WordNet [6]. The integration is performed via an automatic mapping and by filling in lexical gaps in resource-poor languages with the aid of Machine Translation (MT). The result is an "encyclopedic dictionary" that provides babel synsets, i.e., concepts and named entities lexicalized in many languages and connected with large amounts of semantic relations.

While the LOD is centered around DBPedia [1], the largest "hub" of Linked Data which provides wide coverage of Named Entities, BabelNet focuses both on word senses and on Named Entities in many languages. Therefore, its aim is to provide full lexicographic and encyclopedic coverage. Compared to YAGO [11], BabelNet integrates WordNet and Wikipedia by means of a mapping strategy based on a disambiguation algorithm, and provides additional lexicalizations resulting from the application of MT.

In the next Section we introduce BabelNet and briefly illustrate its features. Then, in Section 3 we provide statistics and in Section 4 we describe the RDF encoding of BabelNet. Finally, we give some conclusions in Section 5.

2 BabelNet

BabelNet [8] encodes knowledge as a labeled directed graph $G = (V, E)$ where V is the set of nodes – i.e., concepts such as balloon and named entities such as Montgolfier brothers – and $E \subseteq V \times R \times V$ is the set of edges connecting

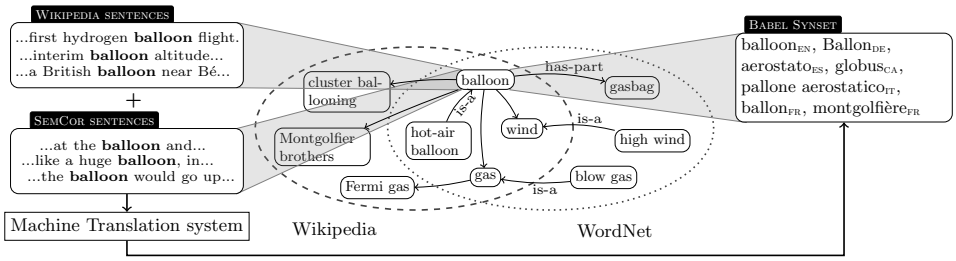

Fig. 2. An overview of BabelNet (nodes are labeled with English lexicalizations only): unlabeled edges are extracted from Wikipages (e.g., Balloon (aircraft) links to Montgolfier brothers), labeled edges come from WordNet (e.g., balloon^1_n *has-part* gasbag^1_n).

pairs of concepts (e.g., balloon *is-a* lighter-than-air craft). Each edge is labeled with a semantic relation from R, e.g., $\{$*is-a*, *part-of*, ..., $\epsilon\}$, where ϵ denotes an unspecified semantic relation. Each node $v \in V$ contains a set of lexicalizations of the concept for different languages, e.g., $\{$ balloon$_{\text{EN}}$, Ballon$_{\text{DE}}$, pallone aerostatico$_{\text{IT}}$, ..., montgolfière$_{\text{FR}}$ $\}$. We call such multilingually lexicalized concepts *Babel synsets*. 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 build the BabelNet graph, we collect at different stages:

a. from WordNet, all available word senses (as concepts) and all the lexical and semantic pointers between synsets (as relations);
b. from Wikipedia, all encyclopedic entries (i.e., Wikipages, as concepts) and semantically unspecified relations from hyperlinked text.

An overview of BabelNet is given in Figure 2. The excerpt highlights that WordNet and Wikipedia can overlap both in terms of concepts and relations: accordingly, in order to provide a *unified resource*, we merge the intersection of these two knowledge sources. Next, 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 **combine WordNet and Wikipedia** by automatically acquiring a mapping between WordNet senses and Wikipages. This avoids duplicate concepts and allows their inventories of concepts to complement each other.
2. We **harvest multilingual lexicalizations** of the available concepts (i.e., Babel synsets) by using (a) the human-generated translations provided by Wikipedia (the so-called *inter-language* links), as well as (b) a machine translation system to translate occurrences of the concepts within sense-tagged corpora.

Language	Lemmas	Synsets	Word senses
English	5,938,324	3,032,406	6,550,579
Catalan	3,518,079	2,214,781	3,777,700
French	3,754,079	2,285,458	4,091,456
German	3,602,447	2,270,159	3,910,485
Italian	3,498,948	2,268,188	3,773,384
Spanish	3,623,734	2,252,632	3,941,039
Total	23,935,611	3,032,406	26,044,643

Table 1. Number of lemmas, synsets and word senses in the 6 languages currently covered by BabelNet.

3. We **establish relations between Babel synsets** by collecting all relations found in WordNet, as well as all wikipedias in the languages of interest: in order to encode the strength of association between synsets, we compute their degree of correlation using a measure of relatedness based on the Dice coefficient.

3 Statistics

In this section we provide statistics for BabelNet 1.0.1, obtained by applying the construction methodology briefly described in the previous Section and detailed in [8].

3.1 WordNet-Wikipedia mapping

The overall mapping contains 89,226 pairs of Wikipages and WordNet senses they map to, covers 52% of the noun senses in WordNet, with an accuracy of about 82% estimated on a random sample of 1,000 items.

3.2 Lexicon

BabelNet currently covers 6 languages, namely: English, Catalan, French, German, Italian and Spanish. Its lexicon includes lemmas which denote both concepts (e.g., balloon) and named entities (e.g., Montgolfier brothers). The second column of Table 1 shows the number of lemmas for each language. The lexicons have the same order of magnitude for the 5 non-English languages, whereas English shows larger numbers due to the lack of inter-language links and annotated sentences for many terms, which prevents our construction approach from providing translations.

In Table 2 we report the number of monosemous and polysemous words divided by part of speech. Given that we work with nominal synsets only, the numbers for verbs, adjectives and adverbs are the same as in WordNet 3.0. As for nouns, we observe a very large number of monosemous words (almost 23 million), but also a large number of polysemous words (more than 1 million).

POS	Monosemous words	Polysemous words
Noun	22,763,265	1,134,857
Verb	6,277	5,252
Adjective	16,503	4,976
Adverb	3,748	733
Total	22,789,793	1,145,818

Table 2. Number of monosemous and polysemous words by part of speech (verbs, adjectives and adverbs are the same as in WordNet 3.0).

		English	Catalan	French	German	Italian	Spanish	Total
English WordNet		206,978	-	-	-	-	-	206,978
Wikipedia	pages	2,955,552	123,101	524,897	506,892	404,153	349,375	4,863,970
	redirections	3,388,049	105,147	617,379	456,977	217,963	404,009	5,189,524
	translations	-	3,445,273	2,844,645	2,841,914	3,046,323	3,083,365	15,261,520
WordNet	monosemous	-	97,327	97,680	97,852	98,089	97,435	488,383
	SemCor	-	6,852	6,855	6,850	6,856	6,855	34,268
Total		6,550,579	3,777,700	4,091,456	3,910,485	3,773,384	3,941,039	26,044,643

Table 3. Composition of Babel synsets: number of synonyms from the English WordNet, Wikipedia pages and translations, as well as translations of WordNet's monosemous words and SemCor's sense annotations.

Both numbers are considerably larger than in WordNet, because – as remarked above – words here denote both concepts (mainly from WordNet) and named entities (mainly from Wikipedia).

3.3 Concepts

BabelNet contains more than 3 million concepts, i.e., Babel synsets, and more than 26 million word senses (regardless of their language). In Table 1 we report the number of synsets covered for each language (third column) and the number of word senses lexicalized in each language (fourth column). 72.3% of the Babel synsets contain lexicalizations in all 6 languages and the overall number of word senses in English is much higher than those in the other languages (owing to the high number of synonyms available in the English WordNet synsets). Each Babel synset contains 8.6 synonyms, i.e., word senses, on average, in any language. The number of synonyms per synset for each language individually ranges from a maximum 2.2 for English to a minimum 1.7 for Italian, with an average of 1.8 synonyms per language.

In Table 3 we show for each language the number of word senses obtained directly from WordNet, Wikipedia pages and redirections, as well as Wikipedia and WordNet translations.

3.4 Relations

We now turn to relations in BabelNet. Relations come either from Wikipedia hyperlinks (in any of the covered languages) or WordNet. All our relations are

	English	Catalan	French	German	Italian	Spanish	Total
WordNet	364,552	-	-	-	-	-	364,552
WordNet glosses	617,785	-	-	-	-	-	617,785
Wikipedia	50,104,884	978,006	5,613,873	5,940,612	3,602,395	3,411,612	69,651,382
Total	51,087,221	978,006	5,613,873	5,940,612	3,602,395	3,411,612	70,633,719

Table 4. Number of lexico-semantic relations harvested from WordNet, WordNet glosses and the 6 wikipedias.

English	WordNet	Large tough nonrigid bag filled with gas or heated air.
	Wikipedia	A balloon is a type of aircraft that remains aloft due to its buoyancy.
German		Ein Ballon ist eine nicht selbsttragende, gasdichte Hülle, die mit Gas gefüllt ist und über keinen Eigenantrieb verfügt.
Italian		Un pallone aerostatico è un tipo di aeromobile, un aerostato che si solleva da terra grazie al principio di Archimede.
Spanish		Un aerostato, o globo aerostático, es una aeronave no propulsada que se sirve del principio de los fluidos de Arquímedes para volar, entendiendo el aire como un fluido.

Table 5. Glosses for the Babel synset referring to the concept of balloon as aircraft'.

semantic, in that they connect Babel synsets (rather than senses), however the relations obtained from Wikipedia are unlabeled.[2] In Table 4 we show the number of lexico-semantic relations from WordNet, WordNet glosses and the 6 wikipedias used in our work. We can see that the major contribution comes from the English Wikipedia (50 million relations) and Wikipedias in other languages (a few million relations, depending on their size in terms of number of articles and links therein).

3.5 Glosses

Each Babel synset naturally comes with one or more glosses (possibly available in many languages). In fact, WordNet provides a textual definition for each English synset, while in Wikipedia a textual definition can be reliably obtained from the first sentence of each Wikipage[3]. Overall, BabelNet includes 4,683,031 glosses (2,985,243 of which are in English). In Table 5 we show the glosses for the Babel synset which refers to the concept of balloon as 'aircraft'.

3.6 Sense-tagged corpus

BabelNet also includes a sense-tagged corpus containing the sentences input to the Machine Translation system. The corpus, called BabelCor, is built by collect-

[2] In a future release of the resource we plan to perform an automatic labeling based on work in the literature. See [7] for recent work on the topic.

[3] "The article should begin with a short declarative sentence, answering two questions for the nonspecialist reader: *What (or who) is the subject?* and *Why is this subject notable?*", extracted from `http://en.wikipedia.org/wiki/Wikipedia:Writing_better_articles`. This simple, albeit powerful, heuristic has been previously used successfully to construct a corpus of definitional sentences [10] and learn a definition and hypernym extraction model [9].

ing from SemCor and Wikipedia those sentences which contain an occurrence of a polysemous word labeled with a WordNet sense (in SemCor) or hyperlinked to a Wikipage (in Wikipedia). A frequency threshold of at least 3 sentences per sense is used in order to make sure that meaningful statistics are computed from the MT system's output, thus ensuring precision. As a result, BabelCor contains almost 2 million sentences (1,986,557 in total, of which 46,155 from SemCor and 1,940,402 from Wikipedia), which provide sense-annotated data for 330,993 senses contained in BabelNet (6,856 from WordNet and 324,137 from Wikipedia).

4 BabelNet in RDF

We now introduce a first RDF encoding of BabelNet. Other encodings, including one in the Lemon RDF model [5], will be made available online soon.

4.1 Babel synsets in RDF

An excerpt of the RDF Babel synset representation follows:

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:bn10schema="http://lcl.uniroma1.it/babelnet/bn10/schema/"
    xmlns:bn10instances="http://lcl.uniroma1.it/babelnet/bn10/instance/">
  ...
  <bn10schema:BabelSynset
    rdf:about="http://lcl.uniroma1.it/babelnet/bn10/instance/bn:00008187n">
    <bn10schema:pos>NOUN</bn10schema:pos>
    <bn10schema:source>WIKIWN</bn10schema:source>
    <bn10schema:babelSynsetId>bn:00008187n</bn10schema:babelSynsetId>
    <bn10schema:mainSense>balloon#n#1</bn10schema:mainSense>
    <bn10schema:semanticallyRelated>
      <bn10schema:BabelSynset
      rdf:about="http://lcl.uniroma1.it/babelnet/bn10/instance/bn:02955250n">
        <bn10schema:mainSense>WIKI:EN:Montgolfier_brothers</bn10schema:mainSense>
      </bn10schema:BabelSynset>
    </bn10schema:semanticallyRelated>
    <bn10schema:hypernym>
      <bn10schema:BabelSynset
      rdf:about="http://lcl.uniroma1.it/babelnet/bn10/instance/bn:00051149n">
        <bn10schema:mainSense>lighter-than-air_craft#n#1</bn10schema:mainSense>
      </bn10schema:BabelSynset>
    </bn10schema:hypernym>
  </bn10schema:BabelSynset>
  <bn10schema:BabelSynset
    rdf:about="http://lcl.uniroma1.it/babelnet/bn10/instance/bn:01631774n">
    <bn10schema:pos>NOUN</bn10schema:pos>
    <bn10schema:source>WIKI</bn10schema:source>
    <bn10schema:babelSynsetId>bn:01631774n</bn10schema:babelSynsetId>
```

```
    <bn10schema:mainSense>WIKI:EN:First_flying_machine</bn10schema:mainSense>
  </bn10schema:BabelSynset>
  <bn10schema:BabelSynset
    rdf:about="http://lcl.uniroma1.it/babelnet/bn10/instance/bn:02955250n">
    <bn10schema:pos>NOUN</bn10schema:pos>
    <bn10schema:source>WIKI</bn10schema:source>
    <bn10schema:babelSynsetId>bn:02955250n</bn10schema:babelSynsetId>
    <bn10schema:mainSense>WIKI:EN:Montgolfier_brothers</bn10schema:mainSense>
  </bn10schema:BabelSynset>
  ...
</rdf:RDF>
```

The excerpt above encodes the three Babel synsets for the concepts of balloon (in the sense of aircraft), first flying machine and Montgolfier brothers. The `<pos>` tag provides the part of speech tag of the synset, the `<source>` tag describes the source from which the synset was obtained (WN for WordNet, WIKI for Wikipedia, WIKIWN for the intersection between the two resources), `<babelSynsetId>` provides the numeric id of the synset, and `<mainSense>` provides the main sense (either from WordNet or Wikipedia) which univocally identifies the Babel synset.

The first Babel synset listed above, i.e., the concept of balloon (`bn:00008187n`), is semantically related to the Montgolfier brothers (`bn:02955250n`), among others, as encoded by the `semanticallyRelated` relation, and is a lighter-than-air craft (`bn:00051149n`), as encoded by the `hypernym` relation.

4.2 Babel senses in RDF

An excerpt of the RDF Babel sense representation follows:

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:bn10schema="http://lcl.uniroma1.it/babelnet/bn10/schema/"
    xmlns:bn10instances="http://lcl.uniroma1.it/babelnet/bn10/instance/">
  ...
  <bn10schema:BabelSense
    rdf:about="http://lcl.uniroma1.it/babelnet/bn10/instance/
                    Balloon_(aircraft)-EN@bn:00008187n">
    <bn10schema:babelSynsetId>bn:00008187n</bn10schema:babelSynsetId>
    <bn10schema:lang>EN</bn10schema:lang>
    <bn10schema:source>WIKI</bn10schema:source>
    <bn10schema:pos>NOUN</bn10schema:pos>
    <bn10schema:lemma>Balloon</bn10schema:lemma>
  </bn10schema:BabelSense>
  <bn10schema:BabelSense
    rdf:about="http://lcl.uniroma1.it/babelnet/bn10/instance/
                    Ballongas-DE@bn:00008187n">
    <bn10schema:babelSynsetId>bn:00008187n</bn10schema:babelSynsetId>
    <bn10schema:lang>DE</bn10schema:lang>
    <bn10schema:source>WIKIRED</bn10schema:source>
```

```
    <bn10schema:pos>NOUN</bn10schema:pos>
    <bn10schema:lemma>Ballongas</bn10schema:lemma>
  </bn10schema:BabelSense>
  <bn10schema:BabelSense
    rdf:about="http://lcl.uniroma1.it/babelnet/bn10/instance/
                    Ballon-DE@bn:00008187n">
    <bn10schema:babelSynsetId>bn:00008187n</bn10schema:babelSynsetId>
    <bn10schema:lang>DE</bn10schema:lang>
    <bn10schema:source>WIKI</bn10schema:source>
    <bn10schema:pos>NOUN</bn10schema:pos>
    <bn10schema:lemma>Ballon</bn10schema:lemma>
  </bn10schema:BabelSense>
  <bn10schema:BabelSense
    rdf:about="http://lcl.uniroma1.it/babelnet/bn10/instance/
                    ballon-FR@bn:00008187n">
    <bn10schema:babelSynsetId>bn:00008187n</bn10schema:babelSynsetId>
    <bn10schema:source>WNTR</bn10schema:source>
    <bn10schema:lang>FR</bn10schema:lang>
    <bn10schema:source>WIKITR</bn10schema:source>
    <bn10schema:pos>NOUN</bn10schema:pos>
    <bn10schema:lemma>ballon</bn10schema:lemma>
  </bn10schema:BabelSense>
  <bn10schema:BabelSense
    rdf:about="http://lcl.uniroma1.it/babelnet/bn10/instance/
              Pallone_aerostatico-IT@bn:00008187n">
    <bn10schema:babelSynsetId>bn:00008187n</bn10schema:babelSynsetId>
    <bn10schema:lang>IT</bn10schema:lang>
    <bn10schema:source>WIKI</bn10schema:source>
    <bn10schema:pos>NOUN</bn10schema:pos>
    <bn10schema:lemma>pallone aerostatico</bn10schema:lemma>
  </bn10schema:BabelSense>
  ...
</rdf:RDF>
```

where `<lang>` represents the language in which the sense is lexicalized, `<source>` is the source from which the sense is obtained (WN for WordNet, WNTR or WIKITR for translations of WordNet- or Wikipedia-annotated text, WIKIRED for a Wikipedia redirection, etc.), `<pos>` is the part of speech tag of the sense, and `<lemma>` specifies the lexicalization for the sense.

5 Conclusions

The Web of Data is in need for multilingual lexicalizations for Linked Open Data. This vision of a Linguistic Linked Open Data (LLOD) has recently been promoted, among others, by the Open Linguistic Working Group as well as other researchers [3]. BabelNet [8] – an ongoing project[4] at the Sapienza Linguistic

[4] Developed in the context of the MultiJEDI ERC Starting Grant: `http://lcl.uniroma1.it/multijedi`.

Computing Laboratory[5] – fits this vision by providing multilingual lexicalizations in RDF for millions of concepts, called Babel synsets, as well as a huge network of semantic relations between them. BabelNet currently covers 6 languages, but is continuously expanded with new information and languages.

Future steps include, among others, the integration of a mapping between BabelNet and other linguistic resources which are already part of the LLOD, such as DBPedia.

Acknowledgments

The author gratefully acknowledges the support of the ERC Starting Grant MultiJEDI No. 259234. The author wishes to thank Giovanni Stilo for his help with the RDF encoding of BabelNet.

References

1. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: Dbpedia - a crystallization point for the web of data. Journal of Web Semantics 7(3), 154–165 (2009)
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. 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)
4. Lassila, O., Swick, R.R.: Resource description framework (rdf) model and syntax specication. In: Technical report, World Wide Web Consortium (1999)
5. McCrae, J., Spohr, D., Cimiano, P.: Linking lexical resources and ontologies on the Semantic Web with Lemon. In: Proceedings of the 8th Extended Semantic Web Conference (ESWC). pp. 245–259. Heraklion, Crete, Greece (2011)
6. Miller, G.A., Beckwith, R., Fellbaum, C.D., Gross, D., Miller, K.: WordNet: an online lexical database. International Journal of Lexicography 3(4), 235–244 (1990)
7. Moro, A., Navigli, R.: WiSeNet: Building a Wikipedia-based semantic network with ontologized relations. In: Proceedings of the 21^{st} ACM Conference on Information and Knowledge Management (CIKM 2012). Maui, HI, USA (2012)
8. 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)
9. Navigli, R., Velardi, P.: Learning Word-Class Lattices for definition and hypernym extraction. In: Proceedings of the 48^{th} Annual Meeting of the Association for Computational Linguistics (ACL). pp. 1318–1327. Uppsala, Sweden (2010)
10. Navigli, R., Velardi, P., Ruiz-Martínez, J.M.: An annotated dataset for extracting definitions and hypernyms from the web. In: Proceedings of the 7th International Conference on Language Resources and Evaluation, Valletta, Malta, 19–21 May 2010 (2010)
11. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A large ontology from Wikipedia and WordNet. Journal of Web Semantics 6(3), 203–217 (2008)

[5] `http://lcl.uniroma1.it`