BabelNet meets Lexicography: the Case of an Automatically-built Multilingual Encyclopedic Dictionary

Roberto Navigli
Sapienza University of Rome
navigli@di.uniroma1.it

Abstract

In this paper we provide a first study of the lexicographic quality of BabelNet, a very large automatically-created multilingual encyclopedic dictionary. BabelNet 2.0, available online at http://babelnet.org, covers 50 languages and provides both lexicographic and encyclopedic knowledge for all the open-class parts of speech. It is obtained from the automatic integration of several language resources, namely: WordNet, Open Multilingual WordNet, Wikipedia and OmegaWiki. Here we present a first analysis of the dictionary entries in terms of their coverage of English and Italian word tokens in a large corpus and in comparison to existing, well-established dictionaries, namely the Oxford Dictionary of English and the Treccani Italian dictionary. We observe that BabelNet contains most meanings of the frequent words under analysis and provides additional, often domain-specific meanings and their textual definitions unavailable in traditional dictionaries, as well as encyclopaedic coverage for those words.

Keywords: Multilinguality; Encyclopedic dictionaries; Quality evaluation of automatically-created dictionaries

1 Introduction

The textual content that is available on the Web is becoming ever increasingly multilingual, providing an additional wealth of valuable information. Most of this information, however, remains inaccessible to the majority of users because of language barriers. Consequently, both humans and automatic systems need tools which will enable them to enjoy the beauty and the usefulness of this varied multilingual world.

The wide majority of bilingual paper dictionaries, however, focus on a given language pair, which are the languages on which the lexicographers, and authors of the dictionary, are expert in. As a result, the sense inventories of dictionaries for different language pairs are different, even if the dictionaries are printed by the same publisher. Integrating these inventories, thereby enabling the creation of a multilingual dictionary, is therefore a very arduous task.

MultiJEDI (Multilingual Joint word sensE Disambiguation, http://multijedi.org) is a major project under way in the Linguistic Computing Laboratory at the Sapienza University of Rome. MultiJEDI is a 5-year Starting Independent Research Grant funded by the European Research Council (ERC) that started in February 2011. The project aims to investigate new, groundbreaking directions in the field of Word Sense Disambiguation (WSD), the task of computationally determining the meaning of words in context (Navigli, 2009; 2012). The key intuition underlying the project is that we now have the capabilities to transform multilinguality from an obstacle to Natural Language Understanding into a powerful catalyst for the task. As a core tool for enabling multilinguality the project aims to create a very large automatically-created multilingual encyclopedic dictionary, called BabelNet, made available online at http://babelnet.org. BabelNet is a novel language resource in several respects, including: being a multilingual dictionary which covers tens of languages; providing both encyclopaedic and lexicographic coverage; including information which is usually not available within dictionaries, such as images, fine-grained category information, multiple textual definitions for the same entry, hyperlinks to other entries, and much more.

Since integrating dictionaries of different kinds and nature, especially on a multilingual scale, is admittedly a hard, ambitious task, in this paper we analyze the lexicographic quality of BabelNet, especially in terms of the user perspective, and compare it against manually created dictionaries, so as to determine the added value of an automatic dictionary integration process. Our analysis is performed both at the corpus level, by studying the coverage provided by BabelNet of word occurrences within text (on a portion of the American National Corpus – ANC), and at the inventory level, i.e. by comparing the BabelNet sense inventory with that of other well-established resources, such as the Oxford Dictionary of English and the Treccani dictionary of Italian. Our analysis shows that the richness and amount of information available in BabelNet largely exceeds that of manually created lexicographic resources.

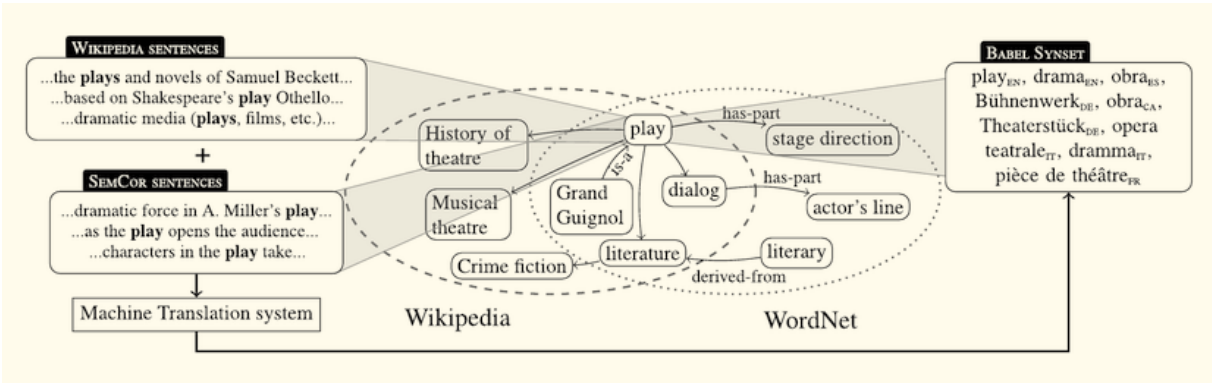

Figure 1: The BabelNet structure.

2 BabelNet 2.0

BabelNet is based on the key idea that different language resources, such as WordNet (Fellbaum, 1998), i.e., the largest machine-readable computational lexicon of English, and Wikipedia (http://www.wikipedia.org), i.e., the most popular multilingual encyclopedia, provide complementary knowledge that can be integrated into a single unified multilingual semantic network covering as many languages as possible. BabelNet, available online at http://babelnet.org, is therefore a large-scale "encyclopedic dictionary". BabelNet 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_{EN}, Theaterstuck_{DE}, dramma_{IT}, obra_{ES}, ..., $\text{piece de theatre}_{FR}$ }. 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 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.

A graphical overview of BabelNet is given in Figure 1. As can be seen, 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 of concepts 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.

Its current version, i.e., BabelNet 2.0, covers 50 languages and provides both lexicographic and encyclopedic knowledge for all the open-class parts of speech. It is obtained from the automatic integration of the following resources:

- WordNet, a popular computational lexicon of English (http://wordnet.princeton.edu, version 3.0);
- Open Multilingual WordNet (http://www.casta-net.jp/~kuribayashi/multi/), a collection of wordnets available in different languages;
- Wikipedia, the largest collaborative multilingual Web encyclopedia (http://wikipedia.org);
- OmegaWiki, a large collaborative multilingual dictionary (http://omegawiki.org).

The number of lemmas for each language ranges between more than 8 million (English) and almost 100,000 (Latvian), with a dozen languages having more than 1 million lemmas. The number of polysemous terms ranges between almost 250,000 in English to only a few thousand for languages such as Galician, Latvian and Esperanto, with most languages having several tens of thousands of polysemous terms. BabelNet 2.0 contains about 9.3 million concepts, i.e., Babel synsets, and above 50 million word senses (regardless of their language). It also contains about 7.7 million images and almost 18 million textual definitions, i.e., glosses, for its Babel synsets. The synsets are linked to each other by a total of about 262 million semantic relations (mostly from Wikipedia). Language distribution of lemmas, synsets and senses in graphically shown in Figure 2. It can be seen that the top 9 languages cover approximately half of the language resource in all respects.

Details on the automatic construction procedure can be found in (Navigli and Ponzetto, 2012) and in (Navigli, 2014), where many applications to Word Sense Disambiguation, Open Information Extraction and Linked Open Data are also reported.

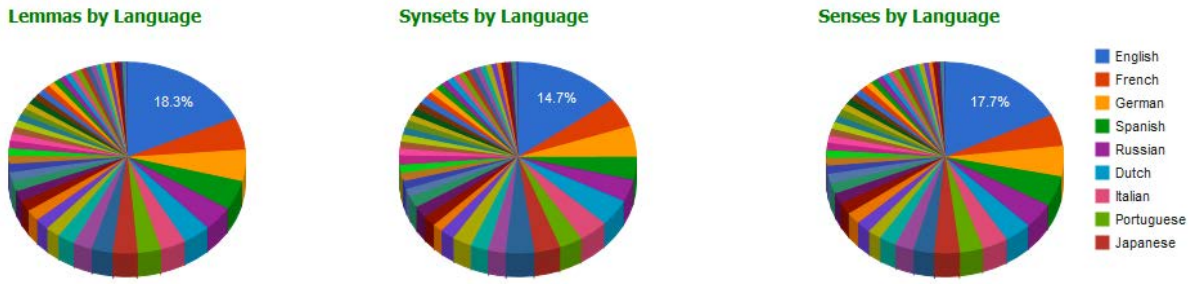

Figure 2: Statistics on the number of lemmas, synsets and senses for the main languages in BabelNet.

3 Corpus coverage in English

To determine corpus coverage, we used the Manually Annotated Sub-Corpus (MASC) (Ide et al., 2008) which consists of parts of the American National Corpus (http://www.anc.org) covering a wide range

of genres of written and spoken textual data amounting to over 500k words. This project aims at organizing and addressing the problems arising against the creation of a resource with multiple annotations. The corpus is available in different formats such as GrAF, in-line XML, token/part of speech sequences, RDF encoding and CoNLL format. The key feature of this corpus is the availability within a single resource of many different linguistic annotations; to date, it contains 17 different types of linguistic annotation, such as sentence boundary, part of speech and syntactic dependency among others. These annotations are the result of a semi-automatic effort in which automatic systems have been coupled with an iterative process of manual evaluations and annotations for retraining the automatic approaches and fine-tuning annotator guidelines to improve inter-annotator agreement.
Moreover, the fact that it is freely available (http://www.anc.org/data/masc/) makes it an invaluable resource for both industry and academic communities in order to produce and improve cutting-edge language technologies.
For our statistics, we considered the set of open-class words in MASC 3.0, totaling 233115 open-class word tokens, and determined, first, the percentage of word tokens for which BabelNet contains an entry for the corresponding lemma and part of speech tag and, second, the percentage of word tokens for which BabelNet contains either a single-word entry or a multiword expression which covers two or more word token in the given sentence. We calculated that 95.15% of open-class word tokens in MASC are covered in BabelNet in the first case, while if we also consider multiword expressions, our coverage increases to 95.53%. We performed the same calculations using the lexicon of the Oxford Dictionary of English (ODE, Soanes & Stevenson, 2003), obtaining 83.91% of single-word tokens covered and 84.03% of tokens covered by any multiword or single-word expression. This shows higher lexicographic coverage (+10%) in BabelNet than in the ODE for the English language. We note that, for many of the uncovered word tokens, the problem is a wrong part-of-speech tag assigned to them (e.g., *achievable* tagged as a noun, *calculus* as an adjective, etc.).
In the future we plan to obtain similar statistics for other languages. However, we note that this requires part-of-speech tagging systems in order to find the appropriate lemma within the dictionary.

4 Dictionary comparison

We performed a comparison of BabelNet against important dictionaries for two different languages, namely: the Oxford Dictionary of English for the English language and the Treccani dictionary for the Italian language.

4.1 English dictionary comparison

As regards English, we compared the lexicographic entries in BabelNet against those of the Oxford Dictionary of English (http://www.oxforddictionaries.com/) for ten of the 1000 most frequent English

lemmas, namely: *work*, *time*, *country*, *head*, *room* (nouns), *remember*, *wait*, *close*, *write*, *contain* (verbs). An analysis of the definitions in the two resources resulted in the following findings:

- **Sense coverage:** in general, the two dictionaries share most of the senses, with additional senses on both sides. However, BabelNet provides a considerably higher number of senses, especially domain-specific ones for nouns and more fine-grained verb sense distinctions. Examples include: a specific thermodynamics sense of *work*, the computer system sense of *time* as well as its representation in ISO time format, *country* in the music style sense, several meanings of *head*, among which: the tip of an abscess, the front a military formation, a difficult juncture and many others; *write* in the sense of coding a computer program. The ODE also includes a few senses which are not covered in BabelNet. For instance, *work* as the operative part of a clock or a defensive structure and *write* in the sense of underwrite (an insurance policy). Finally, we note that BabelNet covers all the most important encyclopaedic meanings of the nominal lemmas, e.g., *head* as the linux program, several films, companies, albums and songs named *Work*, *Time*, *Country* and so on.
- **Quality of sense definitions:** the quality of the sense definitions in the Oxford Dictionary of English is generally higher, with carefully selected usage examples. BabelNet, however, has the advantage of providing several synonyms for the same word sense (e.g., caput, mind, brain, psyche, chief, head word etc. for different meanings of *head*, piece of work, employment, study, mechanical work for *work*, etc.)
- **Quantity of sense definitions:** The number of definitions per sense is considerably higher in BabelNet, thanks to its integration of different language resources. We show statistics in Table 1 (left). It can be seen that, for nouns, BabelNet provides five times the number of definitions per lemma on average while, for verbs, this difference drops to less than 3 times, which is still very high. Interestingly, for nouns BabelNet provides several multiple definitions for the same sense.

4.2 Italian dictionary comparison

We then compared the quality of ten of the 1000 most frequent Italian lemmas in BabelNet against the Treccani Italian dictionary (http://www.treccani.it/vocabolario), namely: *lavoro*, *tempo*, *paese*, *testa*, *sala* (nouns), *ricordare*, *aspettare*, *chiudere*, *scrivere*, *contenere* (verbs). An analysis of the definitions resulted in the following findings:

- **Sense coverage:** in general, the two dictionaries share most of the senses, with additional senses on both sides. Like for English, BabelNet provides coverage for very domain-specific nominal senses, such as *work* in project management, *work* in applied sciences, the linguistic sense of *tempo*, *testa* as the word in a grammatical constituent; the Treccani dictionary, instead, tends to encode all the traditional, regional or historical lexicographic sense distinctions of our words, including some which – due to lack of translations into Italian – are unavailable in BabelNet. Examples include: *sala* in the sense of the complex of acts by which a change of ownership was made in Ger-

manic law; *testa* in the regional Apulian use denoting a species of fish, i.e., *Trigla*; an uncommon usage of *paese* as painted landscape (as in *pittore di paesi*). As regards verbs, we did not find relevant differences between the two dictionaries. Finally, we note that BabelNet covers all the most important encyclopaedic meanings of the nominal lemmas, including a town in Italy called *Paese*, a magazine and a company producing tissues called *Tempo*, several towns and a necropolis called *Sala*, a surname and a novel called *Testa*, etc.

- **Quality of sense definitions:** the quality of the sense definitions in the Treccani dictionary is generally higher, with carefully selected usage examples. However, BabelNet has the big advantage of providing several synonyms for the same word sense (e.g. *opera* for the piece of work sense of *lavoro*; *collocamento, impiego* and *occupazione* for its employment sense, *compito, faccenda, incarico* and *incombente* for its undertaking sense, etc.).
- **Quantity of sense definitions:** The number of definitions per sense is considerably higher in BabelNet, thanks to its integration of different language resources. We show statistics in Table 1 (right). It can be seen that we have a considerably lower number of sense definitions in BabelNet. This is due to the fact that many of the lexical resources integrated, while providing much lexicographic coverage, do not provide textual definitions for the senses they encode. This is particularly true for verbs (and adjectives and adverbs), to which resources like Wikipedia cannot contribute. Interestingly, however, BabelNet provides a higher number, more than twice overall, of senses than the Treccani dictionary, thanks to its integration of several different language resources contributing to its lexical richness also in non-English languages.

		English		Italian	
		BabelNet	ODE	BabelNet	Treccani
Nouns	Total (average) # of senses	79 (15.8)	29 (5.8)	82 (16.4)	30 (6.0)
	Total (average) # of definitions	126 (25.2)	29 (5.8)	37 (7.4)	93 (18.6)
Verbs	Total (average) # of senses	45 (9.0)	17 (3.4)	35 (7.0)	19 (3.8)
	Total (average) # of definitions	50 (10.0)	17 (3.4)	3 (0.6)	44 (8.8)
Total	Total (average) # of senses	124 (12.4)	46 (4.6)	117 (11.7)	49 (4.9)
	Total (average) # of definitions	176 (17.6)	46 (4.6)	40 (4.0)	137 (13.7)

Table 1: Statistics of our ten frequent words for English (left) and Italian (right) using two different dictionaries. Only lexicographic entries are considered (BabelNet encyclopaedic synsets are excluded from these statistics).

4.3 Validation of lexicographic entries with Video Games with a Purpose

As BabelNet is the output of an automatic mapping algorithm (Navigli and Ponzetto, 2012), some of the entries which contain information from several resources, e.g. both WordNet and Wikipedia, might have been merged incorrectly starting from two different senses of the same word. Moreover,

the automatic translation system used to increase the set of multilingual lexicalizations of our Babel synsets might produce wrong translations.

We therefore proposed validating BabelNet using video games with a purpose (Vannella et al. 2014). The annotation tasks are transformed into elements of a video game where players perform their task by playing the game, rather than by performing a more traditional annotation task. While prior efforts in Natural Language Processing have incorporated games for performing the annotation and validation task (Siorpaes and Hepp, 2008; Herdagdelen and Baroni, 2012; Poesio et al., 2013), these games have largely been text-based. In contrast, this year we proposed two video games with graphical 2D gameplay, whose fun nature provides an intrinsic motivation for players to keep playing, thereby increasing the quality of their work and keep the cost per annotation low. The first game, Infection, validates concept-concept relations, and the second, The Knowledge Towers, validates image-concept relations. In experiments involving online players, we demonstrated that, first, players do not need financial incentives to increase the quality of their annotations, second, in a comparison with crowdsourcing, we demonstrated that video game-based annotations consistently generated higher-quality annotations and, third, we found that video game-based annotation can be more cost-effective than crowdsourcing or annotation tasks with game-like features. However, these games did not focus on the validation of the lexicographic entry itself, but on hyperlinks between entries and concept-associated images in BabelNet.

In the future we plan to develop video games that will enable the addition, integration and validation of textual definitions, as well as the validation and addition of senses in arbitrary languages.

4.4 General remarks

Our objective was not to show that BabelNet is better than a traditional dictionary, especially for re source-rich languages such as Italian and English. However, our first analysis shows that, thanks to its integration of several online resources, a multilingual dictionary such as BabelNet provides adequate coverage of lexicographic entries while at the same time containing several synonyms, multiple definitions, hyperlinks to other senses in the dictionary, encyclopedic coverage, which is inherently impossible to achieve in a traditional dictionary, and, last but not least, multilingual interlinking across senses.

In our evaluation we have not taken into account many other features of BabelNet, such as its semantic network structure, which can be explored by humans to better understand the semantics of a concept and exploited by machines to perform automatic tasks such as Word Sense Disambiguation and Entity Linking (Moro et al., 2014), and its availability a Linked Open Data (LOD) thanks to a Lemon-RDF encoding of the network (Ehrmann et al., 2014).

5 Conclusion

In this paper we first presented BabelNet, a multilingual encyclopaedic dictionary automatically constructed from online language resources, and then performed a first qualitative analysis of the BabelNet inventory. Our analysis was performed both in terms of coverage of a large English corpus, i.e., MASC, a subset of the American National Corpus, also in comparison with the Oxford Dictionary of English (ODE), and in terms of coverage and quality of the entries when compared to the ODE for English and the Treccani dictionary of Italian on a random sample of 10 frequent words for the two languages.

6 References

Ehrmann, M., Cecconi, F., Vannella, D., McCrae, J., Cimiano, P., Navigli, R. (2014) Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0. Proc. of *the 9th Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland, 26-31 May, 2014

Fellbaum, C. editor. (1998). WordNet: An Electroni Database. MIT Press.

Herdagdelen, A. & Baroni, M. (2012). Bootstrapping a game with a purpose for common sense collection. *ACM Transactions on Intelligent Systems and Technology*, 3(4):1–24.

Ide, N., Baker, C. Fellbaum, C., Fillmore, C. & Passonneau, R. (2008). MASC: the Manually Annotated Sub-Corpus of American English. In *Proceedings of LREC 2008*.

Moro, A., Raganato, A. & Navigli, R. (2014). Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics* (TACL).

Navigli, R. (2009). Word Sense Disambiguation: a Survey. *ACM Computing Surveys*, 41(2), ACM Press, 2009, pp. 1-69.

Navigli, R. (2012). A Quick Tour of Word Sense Disambiguation, Induction and Related Approaches. In *Proc. of the 38th Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM 2012)*, Spindleruv Mlyn, Czech Republic, January 21-27th, 2012, pp. 115-129.

Navigli, R. & Ponzetto, S. P. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193, Elsevier, 2012, pp. 217-250.

Navigli, R. (2014) BabelNet and Friends: A manifesto for multilingual semantic processing. *Intelligenza Artificiale*, 7(2), pp. 165-181, IOS Press, 2013.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., Ducceschi, L. (2013). Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Interactive Intelligent Systems*, 3(1):3:1–3:44, April.

Soanes, C. & A. Stevenson, editors (2003). Oxford Dictionary of English. Oxford University Press.

Siorpaes, K. & Hepp, M. (2008) Ontogame: Weaving the Semantic Web by online games. In Sean Bechhofer, Manfred Hauswirth, Jrg Hoffmann, and Manolis Koubarakis, editors, *The Semantic Web: Research and Applications, vol. 5021 of Lecture Notes in Computer Science*, pp. 751-766. Springer Berling, Heidelberg.

Vannella, D., Jurgens, D., Scarfini, D., Toscani, D., Navigli, R. (2014) Validating and Extending Semantic Knowledge Bases using Video Games with a Purpose. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, Baltimore, USA, June 22-27, 2014.

Acknowledgements

The author gratefully acknowledges the support of the ERC Starting Grant MultiJEDI No. 259234. Additional thanks go to Daniele Vannella and Andrea Moro for their help with calculating the BabelNet statistics.