Automatic Identification and Disambiguation of Concepts and Named Entities in the Multilingual Wikipedia

Federico Scozzafava(✉), Alessandro Raganato, Andrea Moro, and Roberto Navigli

Dipartimento di Informatica, Sapienza Università di Roma,
Viale Regina Elena 295, 00161 Roma, Italy
federico.scozzafava@gmail.com,
{raganato,moro,navigli}@di.uniroma1.it

Abstract. In this paper we present an automatic multilingual annotation of the Wikipedia dumps in two languages, with both word senses (i.e. concepts) and named entities. We use Babelfy 1.0, a state-of-the-art multilingual Word Sense Disambiguation and Entity Linking system. As its reference inventory, Babelfy draws upon BabelNet 3.0, a very large multilingual encyclopedic dictionary and semantic network which connects concepts and named entities in 271 languages from different inventories, such as WordNet, Open Multilingual WordNet, Wikipedia, OmegaWiki, Wiktionary and Wikidata. In addition, we perform both an automatic evaluation of the dataset and a language-specific statistical analysis. In detail, we investigate the word sense distributions by part-of-speech and language, together with the similarity of the annotated entities and concepts for a random sample of interlinked Wikipedia pages in different languages. The annotated corpora are available at http://lcl.uniroma1.it/babelfied-wikipedia/.

Keywords: Semantic annotation · Named entities · Word senses · Disambiguation · Multilinguality · Corpus annotation · Sense annotation · Word sense disambiguation · Entity linking

1 Introduction

The exponential growth of the Web has resulted in an increased number of Internet users of diverse mother-tongues, and textual information available online in a wide variety different languages. This has led to a heightened interest in multilingualism [9], [15]. Over the last decade, collaborative resources like Wikipedia (an online encyclopedia) and Wiktionary (an online dictionary) have grown not only quantitatively, but also in terms of their degree of multilingualism, i.e., the range of different languages in which they are available. For this reason these resources have been exploited in many Natural Language Processing tasks, such as Word Sense Disambiguation (WSD) [16], [22], [26], i.e., the task of determining the sense of a word in a given context, and Entity Linking (EL) [30], i.e.,

M. Gavanelli et al. (Eds.): AI*IA 2015, LNAI 9336, pp. 1–10, 2015.
DOI: 10.1007/978-3-319-24309-2_27

the task of discovering which named entities are mentioned in a text. Although there are knowledge bases that incorporate these different kinds of knowledge [25], i.e. encyclopedic and lexicographic knowledge, currently there are only few datasets that integrate annotations from both kinds of repositories. This is due to the fact that the research community has typically focused its attention on WSD and EL tasks separately. The main difference between WSD and EL lies in the kind of inventory used, i.e., dictionary vs. encyclopedia respectively; however the tasks are pretty similar, as they both involve the disambiguation of textual mentions according to a reference inventory. Recently, work in the direction of joint word sense and named entity disambiguation has been promoted in order to concentrate research efforts on the common aspects of the two tasks, such as identifying the right meaning in context [21]. The system presented by [21], called Babelfy, attains state-of-the art accuracy in both WSD and EL tasks, including in a multilingual setting. As its sense inventory Babelfy draws upon BabelNet, a multilingual encyclopedic dictionary, that has lexicographic and encyclopedic coverage of terms.

Moreover, a first corpus annotated with both concepts and named entities has also been created [20]. However, as this corpus is only in English, we decided to annotate a sample of Wikipedia automatically with both word senses and named entities in two languages.

The paper is organized as follows. In Section 2 we cover related work on annotated text with senses. In Section 3, 4 and 5 we briefly describe Wikipedia, BabelNet and Babelfy respectively. In Section 6 and 7 we provide statistics and evaluations. Finally, Section 8 presents our conclusions.

2 Related Work

Over the years, several datasets annotated either with concepts or with named entities have been created [7], [20], [31]. Moreover, numerous tasks in competitions such as Senseval/Semeval [12–14], [19], [23,24], [27], [29], [32], TAC KBP EL [11], Microposts [1] and ERD [3] have been organized together with the development of frameworks for comparison of entity-annotation systems [4], [34].

As regards WSD, i.e., the task of determining the sense of a word in a given context, many disambiguation competitions resulted in several datasets that were manually annotated with word senses. However, these datasets are pretty small. In fact, the largest dataset manually annotated with word senses is SemCor [18], a subset of the English Brown Corpus, containing 360K words and more than 200K sense-tagged content words according to the WordNet lexical database [28]. However, these datasets contain only lexicographic annotations without considering named entities, and moreover no dataset of comparable size to SemCor exists for languages other than English.

The EL task, i.e., the task of discovering which named entities are mentioned, was introduced more recently [30]. Usually the reference inventory for this task is Wikipedia. In fact, the largest manually annotated dataset is Wikilinks [31], which contains links to Wikipedia pages. Wikilinks consists of web pages,

crawled from Google's web index, containing at least one hyperlink that points to English Wikipedia. This dataset consists of roughly 13M documents with 59M annotated mentions. Another well-known corpus is the Freebase Annotations of the ClueWeb Corpora (FACC1) [7], built by researchers at Google, who annotated English-language Web pages from the ClueWeb09 and ClueWeb12 corpora. The corpus consists of an automatic annotation of 800M documents with 11 billion entity annotations. The annotations are generally of high quality with a precision around 80-85% and a recall around 70-85% (as stated by the authors). However, as for the WSD task, these resources contain only named entities without taking into account word senses, and are available only for the English language. Recently, [21] proposed a new unified approach for WSD and EL, called Babelfy, which jointly disambiguates word senses and named entities, reaching the state of the art on both tasks in a multilingual setting. Babelfy was used in the first automatic semantic annotation of both named entities and word senses on the MASC corpus [10], [20]. However, this resource is smaller compared to other resources, containing roughly 200K total annotations available only for the English language. In this paper we perform a high quality automatic annotation with both word senses and named entities of a large corpus of English and Italian.

3 Wikipedia

Wikipedia[1] is a well-known freely available collaborative encyclopedia, containing 35 million pages in over 250 languages. The Wikipedia internal links (see Figure 1) are one of the features that makes Wikipedia a valuable project and resource. In fact it was estimated that the network of internal links offers the opportunity to proceed from any article to any other with an average of 4.5 clicks [5].

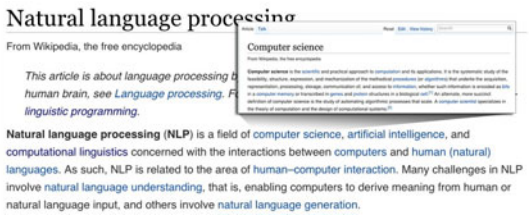

Fig. 1. A sample Wikipedia page with links.

The freedom to create and edit pages has a positive impact both qualitatively and quantitatively, matching and overcoming the famous *Encyclopedia Britannica* [8]. It was estimated that the text of the English Wikipedia is currently equivalent to over 2000 volumes of the *Encyclopedia Britannica*.

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

Wikipedia users are free to create new pages following the guidelines provided by the encyclopedia. In fact, each article in Wikipedia is identified by a unique identifier allowing the creation of shortcuts, expressed as: [[ID |anchor text]], where the anchor text is the fragment of text of a page linked to the identified page ID, and [[anchor text]], where the anchor text is linked to the corresponding homonymous page.

For instance, in the following sentence taken from the Wikipedia page Natural Language Processing: *"Natural language processing (NLP) is a field of [[computer science]], [[artificial intelligence]], and [[computational linguistics]] concerned with the interactions between [[computer]]s and [[Natural language|human (natural) languages]]. As such, NLP is related to the area of [[human-computer interaction]]. Many challenges in NLP involve [[natural language understanding]], that is, enabling computers to derive meaning from human or natural language input, and others involve [[natural language generation]]."*, the users decided to link *human (natural) languages* to the Wikipedia page *Natural language*.

In our settings by exploiting the Babelfy disambiguation system we leverage these hand-made connections to improve the quality of our automatic annotation.

4 BabelNet

Our reference inventory is BabelNet[2] [25], version 3.0, a multilingual lexicalised semantic network obtained from the automatic integration of heterogeneous resources such as WordNet [17], Open Multilingual WordNet [2], Wikipedia[3], OmegaWiki[4], Wiktionary[5] and Wikidata[6]. The integration is performed via an automatic mapping between these resources which results in merging equivalent concepts from the different resources. BabelNet covers and links named entities and concepts present in all the aforementioned resources obtaining a wide coverage resource containing both lexicographic and encyclopedic terms.

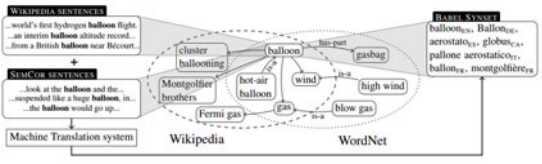

Fig. 2. An illustrative overview of BabelNet (picture from [25]).

[2] http://babelnet.org
[3] http://www.wikipedia.org
[4] http://www.omegawiki.org
[5] http://www.wiktionary.org
[6] http://www.wikidata.org

For instance in Figure 2 the concepts *balloon, wind, hot-air balloon and gas* are defined in both Wikipedia and WordNet while *Montgolfier brothers* and *blow gas* are respectively named entities and concepts retrieved from Wikipedia and WordNet. Each node in BabelNet, called Babel synset, represents a given meaning and contains all the synonyms, glosses and translations harvested from the respective resources. The latest release of BabelNet, i.e., 3.0, provides a full-fledged taxonomy [6], covers 271 languages and it is made up of more than 13M Babel synsets, with 117M senses and 354M lexico-semantic relations (for more statistics see http://babelnet.org/stats). It is also available as SPARQL endpoint and in RDF format containing up to 2 billion RDF triples.

5 Babelfy

To perform the automatic annotation of the considered dataset with both concepts and named entities, we used the latest version of Babelfy[7], i.e., version 1.0. Babelfy is a unified graph-based approach to Entity Linking and Word Sense Disambiguation, a state-of-the-art system in both tasks. A detailed description of the system can be found in [21]. Differently from version 0.9.1, this new release features many parameters among which adding pre-annotated fragments of text to help the disambiguation phase and to enable or disable the most common sense (MCS) backoff strategy that returns the most common sense for the text fragment when the system does not have enough information to select a meaning. Therefore we exploit the links of Wikipedia which are contained in BabelNet as pre-annotated fragments of text.

Babelfy is based on the BabelNet 3.0 semantic network and jointly performs disambiguation and entity linking in three steps. The first step associates with each node of the network a set of semantically relevant vertices, i.e. concepts and named entities, thanks to a notion of semantic signatures. This is a preliminary step which needs to be performed only once, independently of the input text. The second step extracts all the textual mentions from the input text, i.e. substrings of text for which at least one candidate named entity or concept can be found in BabelNet. Consequently, for each extracted mention, it obtains a list of the possible meanings according to the semantic network. The last step consists of connecting the candidate meanings according to the previously-computed semantic signatures. It then extracts a dense sub-graph and selects the best candidate meaning for each fragment.

Therefore our approach is comprised of two main phases: the identification of the semantic context given by the BabelNet synset corresponding to the link in the page (see Figure 3) and the disambiguation of the Wikipedia article through the use of Babelfy. Each Wikipedia page, together with its internal links, corresponds to a Babel synset. Thus providing that information (i.e. the Babel synset) as disambiguation context for the text associated with the link in the page helps the Babelfy algorithm exclude less relevant candidates.

[7] http://babelfy.org

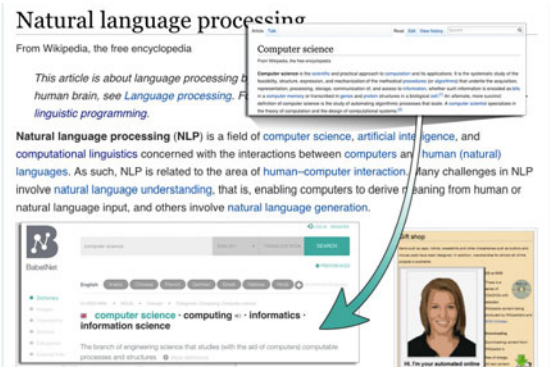

Fig. 3. *Computer science* Wikipedia link with its relative Babel synset.

6 Statistics

In this section we present the statistics of our automatically annotated dataset. We used a sample of 500K articles of English Wikipedia and over 450K articles of Italian Wikipedia POS tagged with the Stanford POS Tagger [33] (for Italian we trained a model using the dataset from the Universal Dependency Treebank Project[8]). The corpora contain respectively 501M and 310M words (see Table 1), among which in both cases 42% are content words (i.e. words POS tagged as noun, adjective, adverb or verb).

Table 1. Statistics of the Wikipedia sample.

	English		Italian	
# Articles		500,000		474,887
# Content Words	209,066,032		133,022,968	
# Non-Content Words	292,796,219		177,786,434	
# Words		501,862,251		310,809,402

In Table 2 and 3, we show the total number of our automatic annotations divided between concepts and named entities with and without the most common sense backoff strategy. As expected we have more annotations with the MCS, while without it we annotated 31% and 21% of the content words, respectively in English and Italian.

7 Evaluation

We performed an evaluation over a restricted sample of annotations to estimate the performance of the system using the accuracy measure, which is defined

[8] https://code.google.com/p/uni-dep-tb/

Table 2. Statistics of our automatic annotation of the Wikipedia corpus with MCS.

	English		Italian	
# Adjective Word Senses	14,662,188		5,921,520	
# Adverb Word Senses	3,402,554		2,604,358	
# Noun Word Senses	55,597,241		31,003,356	
# Verb Word Senses	26,072,320		11,942,285	
# Word Senses		99,734,303		51,471,519
# Named Entities		14,162,561		5,503,556
# Total Number of annotations		113,896,864		56,975,075

Table 3. Statistics of our automatic annotation of the Wikipedia corpus without MCS.

	English		Italian	
# Adjective Word Senses	7,816,765		2,848,886	
# Adverb Word Senses	2,450,533		1,385,650	
# Noun Word Senses	32,398,013		14,313,556	
# Verb Word Senses	8,683,852		3,302,068	
# Word Senses		51,349,163		21,850,160
# Named Entities		14,162,220		5,469,766
# Total Number of annotations		65,511,383		27,319,926

Table 4. Annotations with and without using the internal Wikipedia's links.

	English	Italian
# Articles	1,000	1,000
# Annotations with Wikipedia links	72,142	72,597
# Annotations without Wikipedia links	71,354	71,236

Table 5. Annotations in common between comparable Wikipedia pages in two languages.

	English	Italian
# Articles	1,000	1,000
# Annotations	258,273	107,448
# Annotations in common	23,439	

as the number of correct meanings/entities over the whole number of manually annotated mentions.

For evaluation purposes, we also annotated 1K articles for both languages with and without using the internal Wikipedia links as help to the disambiguation phase. As we can see from Table 4, using the Wikipedia page internal links as semantic context for disambiguation, as described in Section 5, the system yields more annotations.

Moreover, we manually evaluated a random sample of 200 concepts and 200 named entities obtaining an estimated accuracy of 77.8% for word senses and 63.2% for named entities for English, and 78.6% and 66% respectively for Italian.

To estimate the similarity of annotated entities and concepts for cross-lingual interlinked Wikipedia pages, we randomly selected 1K English Wikipedia articles and their equivalent in Italian. In Table 5 we show the number of common annotations (i.e. Babel synsets) between corresponding articles in the two languages.

8 Conclusion

In this paper we presented a large sample of the English and Italian Wikipedias disambiguated with both named entities and concepts, thanks to the use of a state-of-the-art disambiguation and entity linking system, i.e., Babelfy [21]. As sense inventory we used BabelNet 3.0, a multilingual encyclopedic dictionary containing lexicographic and encyclopedic terms obtained from the automatic integration of WordNet, Open Multilingual WordNet, Wikipedia, OmegaWiki, Wiktionary and Wikidata. In order to obtain high quality annotations, we exploited the internal links of Wikipedia as an additional aid for the disambiguation phase. We performed a manual evaluation of our automatic annotation, which indicated an estimated accuracy of 77.8% for word senses, 63.2% for named entities in English, and 78.6% and 66%, respectively, in Italian.

The annotated corpora are available at http://lcl.uniroma1.it/babelfied-wikipedia/.

Acknowledgments

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

References

1. Basave, A.E.C., Rizzo, G., Varga, A., Rowe, M., Stankovic, M., Dadzie, A.S.: Making sense of microposts (#Microposts2014) named entity extraction & linking challenge. In: 4th Workshop on Making Sense of Microposts (#Microposts2014) (2014)
2. Bond, F., Foster, R.: Linking and extending an open multilingual wordnet. In: ACL (1), pp. 1352–1362 (2013)
3. Carmel, D., Chang, M.W., Gabrilovich, E., Hsu, B.J.P., Wang, K.: ERD'14: entity recognition and disambiguation challenge. In: ACM SIGIR Forum, vol. 48, pp. 63–77. ACM (2014)
4. Cornolti, M., Ferragina, P., Ciaramita, M.: A framework for benchmarking entity-annotation systems. In: Proc. of WWW, pp. 249–260 (2013)
5. Dolan, S.: Six Degrees of Wikipedia (2008). http://mu.netsoc.ie/wiki/

6. Flati, T., Vannella, D., Pasini, T., Navigli, R.: Two is bigger (and better) than one: the wikipedia bitaxonomy project. In: Proc. of ACL, pp. 945–955. Association for Computational Linguistics, Baltimore (2014)
7. Gabrilovich, E., Ringgaard, M., Subramanya, A.: FACC1: Freebase annotation of ClueWeb corpora, Version 1. Release date, pp. 06–26 (2013)
8. Giles, J.: Internet encyclopaedias go head to head. Nature **438**(7070), 900–901 (2005)
9. Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., McCrae, J.: Challenges for the multilingual web of data. Web Semantics: Science, Services and Agents on the World Wide Web **11**, 63–71 (2012)
10. Ide, N., Baker, C., Fellbaum, C., Fillmore, C.: MASC: the manually annotated sub-corpus of American English. In: Proc. of LREC (2008)
11. Ji, H., Dang, H., Nothman, J., Hachey, B.: Overview of tac-kbp2014 entity discovery and linking tasks. In: Proc. of TAC (2014)
12. Lefever, E., Hoste, V.: Semeval-2010 task 3: cross-lingual word sense disambiguation. In: Proc. of SemEval, pp. 15–20 (2010)
13. Lefever, E., Hoste, V.: Semeval-2013 task 10: cross-lingual word sense disambiguation. In: Proc. of SemEval, pp. 158–166 (2013)
14. Manandhar, S., Klapaftis, I.P., Dligach, D., Pradhan, S.S.: SemEval-2010 task 14: word sense induction & disambiguation. In: Proc. of SemEval, pp. 63–68 (2010)
15. McDonald, R.T., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K.B., Petrov, S., Zhang, H., Täckström, O., et al.: Universal dependency annotation for multilingual parsing. In: ACL (2), pp. 92–97 (2013)
16. Mihalcea, R.: Using wikipedia for automatic word sense disambiguation. In: HLT-NAACL, pp. 196–203 (2007)
17. Miller, G.A.: WordNet: a lexical database for English. Communications of the ACM **38**(11), 39–41 (1995)
18. Miller, G.A., Leacock, C., Tengi, R., Bunker, R.T.: A semantic concordance. In: Proc. of the workshop on Human Language Technology, pp. 303–308 (1993)
19. Moro, A., Navigli, R.: SemEval-2015 task 13: multilingual all-words sense disambiguation and entity linking. In: Proc. of SemEval, pp. 288–297 (2015)
20. Moro, A., Navigli, R., Tucci, F.M., Passonneau, R.J.: Annotating the MASC corpus with BabelNet. In: Proc. of LREC, pp. 4214–4219 (2014)
21. Moro, A., Raganato, A., Navigli, R.: Entity Linking meets Word Sense Disambiguation: A Unified Approach. TACL **2**, 231–244 (2014)
22. Navigli, R.: Word sense disambiguation: A survey. ACM Computing Surveys (CSUR) **41**(2), 10 (2009)
23. Navigli, R., Jurgens, D., Vannella, D.: Semeval-2013 task 12: multilingual word sense disambiguation. In: Proc. of SemEval, vol. 2, pp. 222–231 (2013)
24. Navigli, R., Litkowski, K.C., Hargraves, O.: Semeval-2007 task 07: coarse-grained english all-words task. In: Proc. of SemEval, pp. 30–35 (2007)
25. 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)
26. Navigli, R., Ponzetto, S.P.: Joining forces pays off: multilingual joint word sense disambiguation. In: Proc. of EMNLP, pp. 1399–1410 (2012)
27. Palmer, M., Fellbaum, C., Cotton, S., Delfs, L., Dang, H.T.: English tasks: all-words and verb lexical sample. In: Proc. of the Second International Workshop on Evaluating Word Sense Disambiguation Systems, pp. 21–24 (2001)

28. Pilehvar, M.T., Navigli, R.: A large-scale pseudoword-based evaluation framework for state-of-the-art word sense disambiguation. Computational Linguistics **40**(4), 837–881 (2014)
29. Pradhan, S.S., Loper, E., Dligach, D., Palmer, M.: Semeval-2007 task 17: English lexical sample, SRL and all words. In: Proc. of SemEval, pp. 87–92 (2007)
30. 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. Springer (2013)
31. Singh, S., Subramanya, A., Pereira, F., McCallum, A.: Wikilinks: a large-scale cross-document coreference corpus labeled via links to Wikipedia. University of Massachusetts, Amherst, Tech. Rep. UM-CS-2012-015 (2012)
32. Snyder, B., Palmer, M.: The English all-words task. In: Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, pp. 41–43 (2004)
33. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: HLT-NAACL, vol. 1, pp. 173–180 (2003)
34. Usbeck, R., Röder, M., Ngonga Ngomo, A.C., Baron, C., Both, A., Brümmer, M., Ceccarelli, D., Cornolti, M., Cherix, D., Eickmann, B., Ferragina, P., Lemke, C., Moro, A., Navigli, R., Piccinno, F., Rizzo, G., Sack, H., Speck, R., Troncy, R., Waitelonis, J., Wesemann, L.: GERBIL - general entity annotation benchmark framework. In: Proc. of WWW, pp. 1133–1143