Wikicorpus: A Word-Sense Disambiguated Multilingual Wikipedia Corpus

[1]Samuel Reese, [1]Gemma Boleda, [1]Montse Cuadros, [1]Lluís Padró, [2]German Rigau

[1]TALP center, [2]IXA NLP Group
Universitat Politècnica de Catalunya (Barcelona), University of the Basque Country (Donostia)
{sreese,gboleda,cuadros,lluisp}@lsi.upc.es, german.rigau@ehu.es

Abstract

This article presents a new freely available trilingual corpus (Catalan, Spanish, English) that contains large portions of the Wikipedia and has been automatically enriched with linguistic information. To our knowledge, this is the largest such corpus that is freely available to the community: In its present version, it contains over 750 million words. The corpora have been annotated with lemma and part of speech information using the open source library FreeLing. Also, they have been sense annotated with the state of the art Word Sense Disambiguation algorithm UKB. As UKB assigns WordNet senses, and WordNet has been aligned across languages via the InterLingual Index, this sort of annotation opens the way to massive explorations in lexical semantics that were not possible before. We present a first attempt at creating a trilingual lexical resource from the sense-tagged Wikipedia corpora, namely, *WikiNet*. Moreover, we present two by-products of the project that are of use for the NLP community: An open source Java-based parser for Wikipedia pages developed for the construction of the corpus, and the integration of the WSD algorithm UKB in FreeLing.

1. Introduction

From its conception, Wikipedia has attracted much interest from researchers in different fields (Medelyan et al., 2008), and it is particularly attractive for Natural Language Processing: It contains hundreds of millions of words of text, reasonably edited, as well as information that enriches this text and can be exploited for NLP purposes (from categories labeling each document to boxes with semi-structured facts about the world). Moreover, its license allows for the use and redistribution of the texts, provided the derived works keep the same GNU Free Documentation License (FDL).[1]

We built a trilingual Wikicorpus with two goals in mind: (1) to build a large, freely available corpus enriched with linguistic information, and (2) to explore the induction of multilingual lexical semantic resources. Wikipedia is a suitable source for texts for both goals, as it has comparable material in different languages. The resulting corpus is by no means balanced (containing only encyclopedic texts), but for our purposes the benefits largely outweigh the drawbacks.

The result of our project is a trilingual corpus (Catalan, Spanish, English) containing over 750 million words, automatically extracted, segmented, lemmatized, part-of-speech tagged, and sense disambiguated with FreeLing (Atserias et al., 2006). It is available for download under FDL at `http://www.lsi.upc.edu/~nlp`, in both tagged and raw text versions. The information in the corpus is complementary to that obtained in related work (Zaragoza et al., 2007; Atserias et al., 2009), and the inclusion of Spanish new (the cited work concerned the English and Catalan Wikipedias, respectively).

To our knowledge, ours is the largest automatically sense-tagged corpus: In its present version, it contains over 750 million words (see Section 3.1.). While much noise can be expected for the WSD output, we believe that the noise will be overweighted by the magnitude of the dataset, so the results of automatic sense assignment are useful for analyzing, evaluating, and developing WSD methods and lexica, as well as for other tasks in NLP and computational linguistics, ranging from Information Extraction to the investigation of linguistic phenomena.

Moreover, two by-products of the project are of use for the NLP community: An open source Java-based parser for Wikipedia pages developed for the construction of the corpus, and the integration of the WSD algorithm UKB (Agirre and Soroa, 2009) in FreeLing.

2. Method

The method consists of three main subprocesses: filtering, text extraction, and linguistic processing. Figure 2. summarizes the process, and the rest of this section provides some details about it.

2.1. Filtering

Not all Wikipedia pages are suitable for corpus construction, for instance redirection pages and desambiguation pages, which typically contain long lists and little text. After some experimentation, we found the best filtering criterion to be to exclude pages without a category[2] (rather than, e.g., the number of in- and out-links). These are typically redirection pages or poorly edited pages.

2.2. Text extraction

The articles, in Mediawiki markup format[3], were accessed using the Java-based Wikipedia Library, JWPL (Zesch et al., 2008). For further processing, we had to extract raw text from the articles. The text output produced by JWPL, however, contained too much noise for its inclusion in a corpus (brackets, file names, etc.). Most alternative parsers[4] convert the markup format to HTML, PDF, or XML, and to adapt them would have required a considerable amount of

[1]http://www.fsf.org/licensing/licenses/fdl.html

[2]Wikipedia pages are assigned one category by human editors, who must not adhere to a preexisting category inventory.

[3]http://www.mediawiki.org/wiki/MediaWiki

[4]http://www.mediawiki.org/wiki/Alternative_parsers

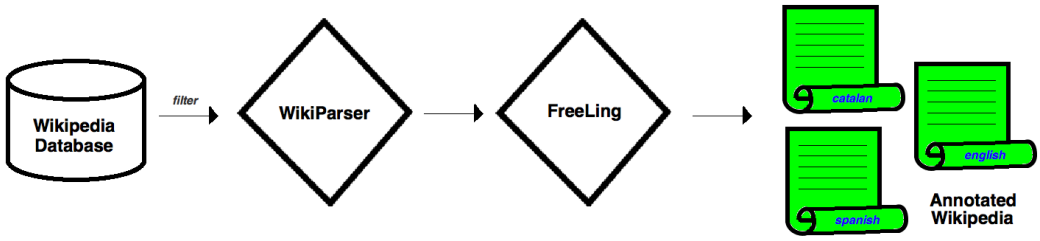

Figure 1: From Wikipedia to an annotated corpus.

Language	Number of lines	Tokens/document
Catalan	49,862,548	344.44
English	118,737,910	447.4
Spanish	611,100,547	455.4

Table 1: Size of the corpora created in the project. The number of lines roughly corresponds to the number of words in the articles. The last column indicates the average number of tokens per document.

effort and development time. For these reasons, we developed a JavaCC parser. While it has some problems (it is not tolerant to errors such as missing brackets), it extracts a high quality output and is relatively fast.

2.3. Linguistic processing

As mentioned above, we used FreeLing for the linguistic processing of the texts in all the languages involved (English, Spanish, and Catalan). The texts were tokenized, lemmatized, and part-of-speech tagged. For our purposes (the construction of a semantic resource), word sense assignment was a requirement. We integrated the graph-based WSD algorithm UKB (Agirre and Soroa, 2009) into FreeLing so we could label the corpus with WordNet senses (Fellbaum, 1998).[5]

3. Results

3.1. Corpus

Table 1 shows the size of the resulting three corpora in their current version (v1.0). The Catalan corpus contains the whole text of the Wikipedia. Due to the problems with the parser mentioned in Section 2.2., some parts of the Spanish and English Wikipedia are still missing (we estimate that around 100 million words are missing for English, and around 80 million words for Spanish).

Figure 2 contains a sample of the corpus, an article from the Spanish Wikipedia, showing the header (label `doc`) and the information (lemma, POS, WordNet sense) added by the use of FreeLing. Note that there is still some noise in the extracted text (semicolon in the second line of the figure) and that only words present in the Spanish WordNet (in this fragment, *regular*) are assigned a sense.

3.2. WikiNet

As mentioned in the introduction, the goal that set us working on the Wikipedia corpus was the construction of a general purpose multilingual lexical semantic resource. Using large-scale knowledge bases, such as WordNet (Fellbaum, 1998), has become a usual, often necessary, practice for most current Natural Language Processing (NLP) systems. Even now, building large and rich enough knowledge bases for broad–coverage semantic processing takes a great deal of expensive manual effort involving large research groups during long periods of development. For example, in more than ten years of manual construction (from 1995 to 2006, that is from version 1.5 to 3.0), WordNet grew from 103,445 to 235,402 semantic relations[6]. This fact has severely hampered the state-of-the-art of current Natural Language Processing (NLP) applications. In fact, hundreds of person-years have been invested in the development of wordnets for various languages (Vossen, 1998), but the data in these resources seems not to be rich enough to support advanced concept-based NLP applications directly. It seems that applications will not scale up to working in open domains without more detailed and rich general-purpose (and also domain-specific) linguistic knowledge built by automatic means.

Fortunately, during the last years the research community has devised a large set of innovative methods and tools for large-scale automatic acquisition of lexical knowledge from structured and unstructured corpora (Lin, 1998; Mihalcea and Moldovan, 2001; Agirre and Martínez, 2001; McCarthy, 2001; Agirre and de la Calle, 2004; Cuadros et al., 2005; Cuadros and Rigau, 2008). Obviously, all these semantic resources have been acquired using a very different set of processes, tools and corpora, resulting on a different set of new semantic relations between synsets. In fact, each semantic resource has different volume and accuracy figures when evaluated in a common and controlled framework (Cuadros and Rigau, 2007). The Wikipedia corpora allow us to test different methods for the automatic acquisition of semantic knowledge (a) at a large scale, (b) on a multilingual dimension.

[5] Currently, the version of WordNet that is integrated in FreeLing (the one we use) is 1.6.

[6] Symmetric relations are counted only once.

```
<doc id="267233" title="Deltoide" dbindex="85004">
; ; Fx 0
En en NP00000 0
geometría geometría NCFS000 0
, , Fc 0
un uno DI0MS0 0
deltoide deltoide NCFS000 0
es ser VSIP3S0 01775973
un uno DI0MS0 0
cuadrilátero cuadrilátero NCMS000 0
no no RN 0
regular regular AQ0CS0 01891762
...
</doc>
```

Figure 2: Sample of the Spanish corpus. The first line contains the header, with information about the document (title, index in the database, etc.). The remaining lines contain the text: one line per word, and one column per linguistic information (word, lemma, part of speech, and WordNet sense).

	Catalan	English	Spanish	More than one
WikiNet5	16,395	32,880	21,545	345
WikiNet10	32,790	65,506	43,090	988
WikiNet15	49,185	97,690	64,635	1,792
WikiNet20	65,580	129,257	86,180	2,745

Table 2: Number of synset-synset relations obtained for each language and for more than one of the three languages (last column) in different versions of WikiNet.

A first attempt at creating a large semantic multilingual resource from the Wikipedia corpus is WikiNet (Reese, 2009). WikiNet has been obtained by automatically inducing synset-synset relations through its representation as a semantic space, using SemanticVectors (Widdows and Ferraro, 2008). The resource aims at being a sort of open source WordNet, automatically induced from a corpus. Because words are sense disambiguated, and because the senses in different languages are mapped through EuroWordNet's InterLingual Index (ILI), such a resource is truly multilingual. However, in its present version WikiNet is not yet mature for use, as the overlap in the relations across languages is very small (see Table 2).[7]

4. Conclusion and future work

We have presented a trilingual corpus with Catalan, Spanish, and English texts from Wikipedia. Overall, it contains over 750 million words. The textual corpus has been automatically acquired, segmented, lemmatized, part-of-speech tagged, and sense disambiguated, and is available for use by any person under the same license as the Wikipedia itself. Moreover, two by-products of the project are of use for the NLP community: An open source Java-based parser for Wikipedia pages developed for the construction of the corpus, and the integration of the WSD algorithm UKB in FreeLing.

We are currently working on optimizing some aspects of the process (most notably, the Java parser and some of the parameters of FreeLing) and including the whole English and Spanish Wikipedias, not only a fragment. We also plan to derive more accurate and larger WikiNets from the complete versions of the multilingual corpora obtained from Wikipedia, with a view to compare methods for the acquisition of semantic knowledge and to develop usable, open-source semantic resources.

We have argued that the Wikicorpus is useful for the construction of massive and multilingual lexical resources, as words are implicitly "aligned" at the sense level, through the WordNet InterLingual Index. An initial attempt at building such a resource, the *WikiNet*, has been presented in this article. As mentioned in Section 1., we believe that the noise introduced by the automatic sense labeling will be overweighted by the magnitude of the dataset, so the results of automatic sense assignment can indeed be used for the construction of lexical resources and other purposes; however, this remains to be tested, and indeed evaluation is the first and most pressing issue that should be addressed in future work.

Future work should also address the relationships between the three corpora, profiting from the fact that they are largely parallel. Currently, the documents in the three corpora are not aligned in any way. It would be very useful for NLP purposes if they were aligned at page level, which involves identifying corresponding pages with differing lengths (as we have shown in Table 1, on average, pages in the Catalan Wikipedia are shorter than their Spanish and

[7] WikiNet was constructed with a preliminary version of the Wikicorpus, containing roughly the same amount of data for Catalan and Spanish as the present version, but only 180 million words for English.

English counterparts), as well as pages with no equivalent in the other languages. This type of alignment could be exploited for applications such as Machine Translation, or cross-lingual summarization.

Another type of alignment that would be very useful is alignment at the level of categories. If categories were aligned cross-lingually, the corpora could be used for the automatic induction (or expansion) of ontologies and thesauri, again with a multilingual dimension, which could help distinguish universal properties of the induced resources from accidental or language-dependent ones (Zesch and Gurevych, 2007). Currently, the corpus does not contain information about the category assigned to each document. Including this piece of information remains also for future work.

Acknowledgments

This research has received support from the projects KNOW (TIN2006-15049-C03-03) and KNOW2 (TIN2009-14715-C04-04), from the Spanish Ministry of Science and Innovation, and from the Juan de la Cierva programme (JCI-2007-57-1479) from the Spanish Ministry of Education. We also thank and Eneko Agirre and Aitor Soroa for help in the integration of UKB into FreeLing and two anonymous reviewers for suggestions.

5. References

Eneko Agirre and Oier Lopez de la Calle. 2004. Publicly available topic signatures for all wordnet nominal senses. In *Proceedings of LREC*, Lisbon, Portugal.

Eneko Agirre and David Martínez. 2001. Learning class-to-class selectional preferences. In *Proceedings of CoNLL01*, Toulouse, France.

Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009)*, Athens, Greece.

Jordi Atserias, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluís Padró, and Muntsa Padró. 2006. Freeling 1.3: Syntactic and semantic services in an open-source nlp library. In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, May. ELRA. http://www.lsi.upc.edu/ nlp/freeling.

Jordi Atserias, Carlos Rodríguez, and Teresa Suñol. 2009. Vikipèdia catalana etiquetada semànticament 1.0. In *Jornada del Processament Computacional del Català*.

Montse Cuadros and German Rigau. 2007. Semeval-2007 task 16: Evaluation of wide coverage knowledge resources. In *Fourth International Workshop on Semantic Evaluations (SemEval-2007)*.

Montse Cuadros and German Rigau. 2008. Knownet: using topic signatures acquired from the web for building automatically highly dense knowledge bases. In *COLING*, August.

Montse Cuadros, Lluís Padró, and German Rigau. 2005. Comparing methods for automatic acquisition of topic signatures. In *Proceedings of the International Conference on Recent Advances on Natural Language Processing (RANLP'05)*, September.

Christiane Fellbaum. 1998. *WordNet. An Electronic Lexical Database*. Language, Speech, and Communication. The MIT Press.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of International Conference on Machine Learning*, Madison, Wisconsin.

Diana McCarthy. 2001. *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Aternations, Subcategorization Frames and Selectional Preferences*. Ph.D. thesis, University of Sussex.

Olena Medelyan, Catherline Legg, David Milne, and Ian H. Witten. 2008. Mining meaning from wikipedia. Working Paper.

Rada Mihalcea and Dan Moldovan. 2001. extended wordnet: Progress report. In *NAACL Workshop* WordNet and Other Lexical Resources: Applications, Extensions and Customizations *(NAACL'2001).*, pages 95–100, Pittsburg, PA, USA.

Samuel Reese. 2009. Wikinet: Construction d'une ressource lexico-sémantique multilingue à partir de wikipedia. Master's thesis, Institut Supérieur de l'Aéronautique et de l'Espace ISAE, formation Supaero.

Piek Vossen. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers.

Dominic Widdows and Kathleen Ferraro. 2008. Semantic vectors: a scalable open source package and online technology management application. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, May.

Hugo Zaragoza, Jordi Atserias, M. Ciaramita, and G. Attardi. 2007. Semantically annotated snapshot of the english wikipedia. http://www.yr-bcn.es/semanticWikipedia.

Torsten Zesch and Iryna Gurevych. 2007. Analysis of the Wikipedia category graph for NLP applications. In *Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing*, pages 1–8, Rochester, NY, USA. Association for Computational Linguistics.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Extracting lexical semantic knowledge from wikipedia and wiktionary. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA), May.