Semantically Annotated Snapshot of the English Wikipedia

Jordi Atserias, Hugo Zaragoza, Massimiliano Ciaramita, Giuseppe Attardi

Yahoo! Research Barcelona, U. Pisa, on sabbatical at Yahoo! Research
C/Ocata 1
Barcelona 08003
Spain
{jordi, hugoz, massi}@yahoo-inc.com, attardi@di.unipi.it

Abstract

This paper describes SW1, the first version of a semantically annotated snapshot of the English Wikipedia. In recent years Wikipedia has become a valuable resource for both the Natural Language Processing (NLP) community and the Information Retrieval (IR) community. Although NLP technology for processing Wikipedia already exists, not all researchers and developers have the computational resources to process such a volume of information. Moreover, the use of different versions of Wikipedia processed differently might make it difficult to compare results. The aim of this work is to provide easy access to syntactic and semantic annotations for researchers of both NLP and IR communities by building a reference corpus to homogenize experiments and make results comparable. These resources, a semantically annotated corpus and a "entity containment" derived graph, are licensed under the GNU Free Documentation License and available from http://www.yr-bcn.es/semanticWikipedia.

1. Introduction

Wikipedia[1], the largest electronic encyclopedia, has become a widely used resource for different Natural Language Processing tasks, e.g. Word Sense Disambiguation (Mihalcea, 2007), Semantic Relatedness (Gabrilovich and Markovitch, 2007) or in the Multilingual Question Answering task at Cross-Language Evaluation Forum (CLEF)[2]. In the field of Information Retrieval, INEX[3] started in 2002 organizing a yearly competition among research groups focusing on XML Retrieval (similar to TREC) using an XML-ized snapshot of the Wikipedia (XML Document Mining Challenge) and it still plans to keep using Wikipedia as test collection in the next edition[4].

Wikipedia is also a natural corpus for investigating the properties and the applicability of current Natural Language Processing (NLP) technologies; e.g., the usefulness of NLP for Information Retrieval (IR) tasks. Different versions of Wikipedia can be obtained through various NLP pre-processing steps, making difficult to compare results from studies by different authors. The SW1 Corpus aims to provide a high-quality, by current standards, snapshot of the Wikipedia that can be used as a reference in scientific research in different fields such as NLP and IR.

The SW1 corpus is a snapshot of the English Wikipedia dated from 2006-11-04 processed with a number of public-available NLP tools. In order to build SW1, we started from the XML-ized Wikipedia dump distributed by the University of Amsterdam[5]. This snapshot of the English Wikipedia contains 1,490,688 entries excluding redirects. Redirects are Wikipedia entries that point to another Wikipedia entry. We do not provide those entries as they can be retrieved from the original XML-ized version.

2. Processing

Starting from the XML Wikipedia source we carried out a number of data processing steps:

- **Basic preprocessing**: Stripping the text from the XML tags and dividing the obtained text into sentences and tokens.
- **PoS tagging** was performed using the SuperSense Tagger (Ciaramita and Altun, 2006) (see below Semantic Tagging) trained on the WSJ Penn Tree Bank (Marcus et al., 1994). This tagger has an accuracy of 97.1% on WSJ.
- **Lemmatization**: Lemmatization was carried out using morphologial functions (morph) of the WordNet library (Miller et al., 1993).
- **Dependency parsing**: DeSR, an open source statistical parser[6] (Attardi et al., 2007) trained on the WSJ Penn Treebank, was used to obtain syntactic dependencies, e.g. Subject, Object, Predicate, Modifier, etc. DeSR achieved an accuracy of 85.85% LAS, 86.99% UAS in the CoNLL 2007 English Multilingual shared task.
- **Semantic Tagging**: Several semantic taggers, see below, were used to assign semantics tags to words, i.e. WordNet SuperSenses, Named Entity tags according to both IEER and CoNLL.

2.1. Semantic Tagging

The SuperSense Tagger[7] (Ciaramita and Altun, 2006) was used for semantic tagging. The basic tagger is a first-order Hidden Markov Model trained with a regularized average perceptron algorithm.

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

[2] http://clef-qa.itc.it/

[3] http://inex.is.informatik.uni-duisburg.de

[4] http://inex.is.informatik.uni-duisburg.de/2007/

[5] http://ilps.science.uva.nl/WikiXML/

[6] http://desr.sourceforge.net

[7] Available at http://sourceforge.net/projects/supersensetag/

Features used are the lowercased word, PoS, shape (regular expression simplification), bi/tri-grams of characters from the beginning and ending of each word, several gazetters and triggers from GATE[8].

We trained three semantic taggers on different data sets to annotate the data:

- **WordNet SuperSenses (WNSS)**: This model was trained using the Semcor Corpus (Miller et al., 1993). The WordNet Supersenses are the top 45 categories of the WordNet synset hierarchy used by lexicographers to organize synsets based on syntactic category and logical groupings. The accuracy of this tagger estimated by crossvalidation is about 80% F1.
- **Wall Street Journal (WSJ)**: This model was trained on the BBN Pronoun Coreference and Entity Type Corpus[9] from LDC, which supplements the WSJ Penn TreeBank with annotation for 105 categories, overall more than one million words. The WSJ annotation tagset consist of 12 named entity types (*Person*, *Facility*, *Organization*, *GPE*, *Location*, *Nationality*, *Product*, *Event*, *Work of Art*, *Law*, *Language*, and *Contact-Info*), nine nominal entity types (*Person*, *Facility*, *Organization*, *GPE*, *Product*, *Plant*, *Animal*, *Substance*, *Disease* and *Game*), and seven numeric types (*Date*, *Time*, *Percent*, *Money*, *Quantity*, *Ordinal* and *Cardinal*). Several of these types are further divided into subtypes. The accuracy of this tagger is about 87% F1.
- **WSJCoNLL**: This model was trained on the BBN Pronoun Coreference and Entity Type Corpus where the WSJ labels were converted into the CoNLL 2003 NER tagset (Sang and Muelder, 2003) using a manually created map. The CoNLL tagset is composed of four categories of phrases: person (PER), organization (ORG), location (LOC) and miscellaneous (MISC). The accuracy of this tagger is about 91% F1.

3. Dictionaries & Graphs

SW1 also provides a graph of the connections between a passage/sentence and the named entities it contains, extracted according to the WSJ tagset. Figure 3 illustrates a fragment of this data.

The graph is supplied in a textual representation, where each line contains tab-separated fields, representing the Named Entity, its type, as well as the internal and external passage ids (composed by the file and the Wikipedia id of the entry and the sentence sequence number within the document).

Using this information an “entity containment” graph, that is the occurrence graph of named entities in sentences can be built. The “entity containment” graph connects each passage to all entities present in the passage. This forms a bipartite graph in which the degree of an entity equals its passage frequency. Figure 1 and 2 show the entity-containment graph for a subset of documents which satisfy the query ’Life of Pablo Picasso’.

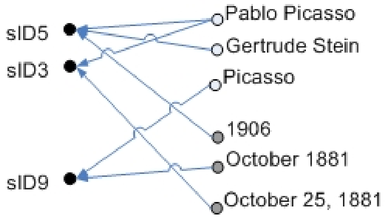

Figure 1: Detailed Graph, ’Life of Pablo Picasso’

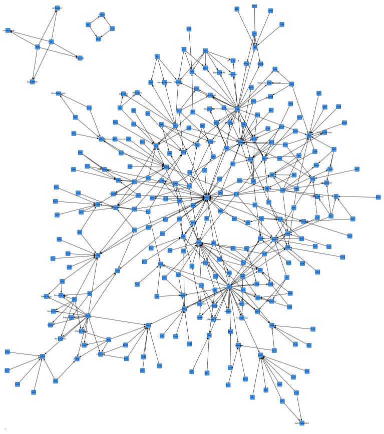

Figure 2: Full Entity Containment Graph

The “entity containment” graph as well as an hyperlink graph of the Wikipedia were built using the open source libraries[10] fastutils, MG4J and WebGraph from the University of Milan. In order two build these graphs two dictionary files listing all the sentence internal ids and all the entities respectively are needed.

4. Distribution

The SW1 snapshot of the Wikipedia contains 1,490,688 entries from which 843,199,595 tokens in 74,924,392 sentences were extracted. Table 1 shows the number of semantics tags for each tagset and the average length in the number of tokens.

	#Tags	Average Length
WNSS	360,499,446	1,27
WSJ	189,655,435	1,70
WSJCoNLL	96,905,672	2,01

Table 1: Semantic Tag figures

The Wikipedia snapshot was divided into 3,000 UTF-8 encoded files each one containing about 500 entries in a format called ’multitag’. The multitag format contains all the Wikipedia text split into sentences plus all the semantic tags. The format is a simple tabular format designed to facilitate the use of the SW1 data.

[8] http://gate.ac.uk/

[9] Linguistic Data Consortium: LDC2005T33

[10] http://vigna.dsi.unimi.it/software.php

DocId:SenId	Named Entity	WSJ Tag	FileId.WikiId.SenId
405750:0	Pablo Picasso	E:PERSON	wiki816.24176.0
405750:1	Picasso	E:WORK_OF_ART:PAINTING	wiki816.24176.1
405750:2	Picasso	E:PERSON	wiki816.24176.2
405750:3	Young Pablo Picasso	E:PERSON	wiki816.24176.3
405750:4	Pablo Picasso	E:PERSON	wiki816.24176.4
405750:4	October 25 , 1881 April 8 , 1973	T:DATE:DATE	wiki816.24176.4
405750:4	Spanish	E:PER_DESC	wiki816.24176.4
405750:4	painter	E:PER_DESC	wiki816.24176.4
405750:4	sculptor	E:PER_DESC	wiki816.24176.4
405750:5	One	N:CARDINAL	wiki816.24176.5
405750:5	20th century	T:DATE:DATE	wiki816.24176.5
405750:5	co-founder	E:PER_DESC	wiki816.24176.5
405750:5	Georges Braque	E:PERSON	wiki816.24176.5
405750:6	Picasso	E:PERSON	wiki816.24176.6

Figure 3: Sentence WSJ Named Entities

%%#DOC wiki816.24176								
%%#PAGE Pablo_Picasso								
.....								
%%#SEN 22476 wx10								
Pablo	NNP	pablo	B-PER	B-noun.person	B-E:PERSON	2	NMOD	0
Picasso	NNP	picasso	I-PER	I-noun.person	I-E:PERSON	14	SBJ	0
(	(	(	0	0	0	4	P	0
October	NNP	october	0	B-noun.time	B-T:DATE:DATE	2	PRN	B-/wiki/October_25
25	CD	25	0	B-adj.all	I-T:DATE:DATE	4	NMOD	I-/wiki/October_25
,	,	,	0	0	I-T:DATE:DATE	4	P	0
1881	CD	1881	0	0	I-T:DATE:DATE	9	NMOD	B-/wiki/1881
	NNP		0	0	I-T:DATE:DATE	9	NMOD	0
April	NNP	april	0	B-noun.time	I-T:DATE:DATE	4	NMOD	B-/wiki/April_8
8	CD	8	0	0	I-T:DATE:DATE	9	NMOD	I-/wiki/April_8
,	,	,	0	0	I-T:DATE:DATE	4	P	0
1973	CD	1973	0	0	I-T:DATE:DATE	4	NMOD	B-/wiki/1973
)	)	)	0	0	0	4	P	0
was	VBD	be	0	B-verb.stative	0	0	ROOT	0
a	DT	a	0	0	0	18	NMOD	0
Spanish	JJ	spanish	B-MISC	B-adj.pert	B-E:NORP:NATIONALITY	18	NMOD	B-/wiki/Spain
painter	NN	painter	0	B-noun.person	B-E:PER_DESC	18	COORD	I-/wiki/Painter
and	CC	and	0	0	0	14	VMOD	0
sculptor	NN	sculptor	0	B-noun.person	B-E:PER_DESC	18	COORD	B-/wiki/Sculpture
.	.	.	0	0	0	14	P	0
%%#SEN 22477 wx11								
One	CD	one	0	0	B-N:CARDINAL	13	ADV	0
of	IN	of	0	0	0	1	NMOD	0
the	DT	the	0	0	0	6	NMOD	0
most	RBS	most	0	B-adv.all	0	5	AMOD	0

Figure 4: Multitag Format Example taken from a fragment of the Pablo Picasso Wikipedia Entry

The multitag format contains one token per line, different attributes associated with each token are separated by tabular characters. Since one tag can span more than one token, the IOB-format is used (*B* Beginnig of the tag, *I* continuation of an open tag, *0* no tag).

Documents are split into sentences and several documents are stored within each file. Lines starting with the sequence *%%#* are used to separate the parts and provide meta information (e.g. information about the document).

Figure 4 illustrates an example of this format, for some sentences of the Wikipedia entry *Pablo Picasso*. The first line encodes the information that this is the Wikipedia entry 24176, the second line provides the title *Pablo_Picasso*. After a sentence mark (*%%#SEN*) each line represent the information associated to each token. The columns for each token are respectively, word form, PoS, lemma, CoNLL, WNSS, WSJ, the head of the dependency relation, its label and links.

For instance, the first token 'Pablo' is tagged as a proper noun (*NNP*) by the PoS tagger and as the beginning of a person (*Pablo Picasso*) according to both the CoNLL tagger (*B-PER*), WSJ tagger (*B-E:PERSON*) and the WNSS tagger (*B-noun.person*). Finally, the syntactic information tell us that 'Pablo' is a noun modifier (*NMOD*) of the token *Picasso* (token number 2) and that there is no link in the original XML containing that token (*0*).

In order to obtain the link information, also represented in IOB-format, the text was realigned with the original XML files. This was possible since the resulting text has almost no modification[11] with respect to the original text in the XMLized version.

Both the semantically annotated corpus and the text representation indicating which sentence is connected to which named entities extracted according to the WSJ are licensed under the GNU Free Documentation License and available from http://www.yr-bcn.es/semanticWikipedia.

5. Usages and Future work

We have presented the first release of a semantically annotated snapshot of the English Wikipedia (SW1) which has been built with the hope it could be valuable resource for both the NLP community and the IR community. This snapshot has been already used in (Zaragoza et al., 2007) and also in the Tag visualiser[12] by Bestiario[13].

Extracting the text to be processed from the XML version not only requires determining which xml tags contain the text to be processed, but also involves segmentation decisions, i.e. which xml tag should split words or sentences. In the current version we have tried to process all the tags that contain text, including metatext, tables, captions, etc.

The nature of the Wikipedia text (especially in the tables, lists, references) differ significally in distributional properties from the corpora used to train the taggers and the parser. Therefore the quality of these NLP processors is considerably lower than that what results from the evaluation in-domain, as has been found before (Ciaramita and Altun, 2005). As a matter of fact we hope than this resource will help understanding the domain adaptation problem for this kind of semantic processing.

As the Wikipedia is growing constantly we may distribute newer versions and explore the new initiatives that appear to better extract the content of the Wikipedia from its original format. For instance, the Wikipedia Extraction (WEX[14]) system which aims to normalize the Wikipedia corpus into a simple, manageable data set.

Another set of extraction and annotation tools has been developed in the project on Semantic Annotation of Italian Wikipedia [15]. These tools include a text extractor that operates directly from a snapshot downloaded from the Wikipedia database as well as a tokenizer, sentence splitter and dependency parser for Italian.

Future releases of SW1 may include not only improved or additional tagsets, but also multilingual versions, since we are planning to create semantically annotated Wikipedia snapshots for other languages.

6. References

G. Attardi, F. Dell'Orletta, M. Simi, A. Chanev, and M. Ciaramita. 2007. Multilingual dependency parsing and domain adaptation using desr. In *Proceedings the CoNLL Shared Task Session of EMNLP-CoNLL 2007*.

M. Ciaramita and Y. Altun. 2005. Named-entity recognition in novel domains with external lexical knowledge. In *Advances in Structured Learning for Text and Speech Processing (NIPS 2005)*.

M. Ciaramita and Y. Altun. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of the EMNLP*.

E. Gabrilovich and S. Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In Manuela M. Veloso, editor, *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1606–1611.

M. Marcus, G. Kim, M. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger. 1994. The Penn treebank: Annotating predicate argument structure.

Rada Mihalcea. 2007. Using wikipedia for automatic word sense disambiguation. In *Proceedings ofThe North American Chapter of the Association for Computational Linguistics(NAACL) HLT 2007*, pages 196–203.

G.A. Miller, C. Leacock, R. Tengi, and R.Bunker. 1993. A semantic concordance. In CA: Morgan Kauf-mann Publishers San Mateo, editor, *Proceedings of the ARPA Human Language Technology Workshop.*, Princeton, NJ.

Erik F. Tjong Kim Sang and Fien De Muelder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *CoNLL 2003 Shared Task*, pages 142–147.

H. Zaragoza, H. Rode, P. Mika, J. Atserias, M. Ciaramita, and G. Attardi. 2007. Ranking very many typed entities on wikipedia. In *CIKM*, pages 1015–1018.

[11] The text is only altered by removing of blank spaces, tabulations, newline and quote normalization

[12] http://www.6pli.org/jProjects/yawibe/

[13] http://www.bestiario.org/web/bestiario.php

[14] http://download.freebase.com/wex/

[15] http://medialab.di.unipi.it/wiki/index.php/Analisi_di_testi_per_il_Semantic_Web_e_il_Question_Answering