Publicly available topic signatures for all WordNet nominal senses

Eneko Agirre, Oier Lopez de Lacalle

IXA NLP Group
University of the Basque Country
{eneko,jibloleo}@si.ehu.es

Abstract

Topic signatures are context vectors built for word senses and concepts. They can be automatically acquired from the web for any concept hierarchy using the "monosemous relative" method. Topic signatures have been shown to be useful in Word Sense Disambiguation, for modeling similarity between word senses, classifying new terms in hierarchies and also building hierarchical clusters of word senses for a given word. In this work we present a publicly available resource which comprises both automatically extracted examples for all WordNet 1.6 noun senses and topic signatures built based on those examples. We gathered around 700 sentences per each noun in WordNet. When the monosemous relatives are used to build a sense corpus for polysemous words, they comprise an average of around 3,500 sentences per word sense. The size of the topic signatures thus constructed is of around 4,500 words per word sense.

1. Introduction

Knowledge acquisition is a long-standing problem in both Artificial Intelligence and Natural Language Processing (NLP). Huge efforts and investments have been made to manually build repositories with semantic and pragmatic knowledge (e.g. EDR, Cyc, Wordnet). Complementary to this, methods to induce and enrich existing repositories have been explored (see (Maedche and Staab, Forthcoming) for a recent review).

In previous work we have shown that it is possible to enrich WordNet synsets with topic signatures. Topic signatures try to associate a topical vector to each word sense. The dimensions of this topical vector are the words in the vocabulary and the weights try to capture the relatedness of the words to the the target word sense. In other words, each word sense is associated with a set of related words with associated weights. For instance, figure 1 shows partially the acquired topic signatures for each sense of *church*. The topic signatures used in this paper can be browsed in full[1], and are publicly available for download[2].

We can build such topic signatures from sense-tagged corpora, just observing which words co-occur distinctively with each sense, or we can try to associate a number of examples from other untagged corpora to each sense and then analyze the occurrences of words in such examples. The words with the most relevant frequencies will constitute the topic signature for each sense.

Topic signatures for words have been successfully used in summarization tasks (Lin and Hovy, 2000). Regarding topic signatures for word senses, Agirre et al.(2000;2001) show that it is possible to obtain good quality topic signatures for word senses automatically. Alfonseca and Manandhar (2002) show that topic signatures for word senses can be used for extending WordNet's taxonomy, Agirre et al. (2004) show that they can be used to compute the similarity between word senses, and Agirre and Lopez (2003) show that they are effective for clustering WordNet word senses. Martinez and Agirre (2004) shows that the method to acquire topic signatures is also relevant for Word Sense Disambiguation (WSD).

This paper is organized as follows. Section 2. explains how to collect the examples and Section 3. presents the method to build topic signatures. Finally, Section 4. presents some conclusions.

2. Method to collect the examples

Corpora where the occurrences of word senses have been manually tagged are a scarce resource. Semcor (Miller et al., 1990) is the largest of all and currently comprises 409,990 word forms. All 190,481 open-class words in the corpus are tagged with word senses. Still, it has a low number of examples for each word sense. The word *bar*, for instance, has 6 word senses, but only 21 occurrences in Semcor.

Other tagged corpora are based on a limited sample of words. For instance, the Senseval-2 English lexical sample corpus comprises 5,266 hand-tagged examples for a set of 29 nouns, yielding an average of 181.3 examples per word. In particular, *bar* has 455 occurrences.

The scarcity of hand-tagged data is the acquisition bottleneck of supervised WSD systems. As an alternative, different methods to build examples for word senses have been proposed in the literature (Leacock et al., 1998; Agirre et al., 2001; Mihalcea, 2002). The methods usually rely on information in WordNet (lexical relations such as synonymy and hypernymy, or words in the gloss) in order to retrieve examples from arge corpora or the web. The retrieved examples might not contain the target word, but they do contain a word that is (closely) related to the target word sense. (Agirre and Martinez, 2004) shows that examples collected following the method described in this paper can be succesfully applied to WSD.

In this work we use WordNet 1.6 as the sense inventory and source of relations. The main reason is compatibility with the MEANING Multilingual Central Repository (Atserias et al., 2004), but we also plan to collect topic signatures for the newer versions.

Before describing the method to collect the examples, we will first define what we mean by "monosemous word".

[1] http://ixa3.si.ehu.es/cgi-bin/signatureak/signaturecgi.cgi

[2] http://ixa2.si.ehu.es/pub/webcorpus

1. sense: church, Christian_church, Christianity "a group of Christians; any group professing Christian doctrine or belief; "

size church(1177.83) catholic(700.28) orthodox(462.17) roman(353.04) religion(252.61) byzantine(229.15) protestant(214.35) rome(212.15) western(169.71) established(161.26) coptic(148.83) jewish(146.82) order(133.23) sect(127.85) old(86.11) greek(68.65) century(61.99) history(50.36) pentecostal(50.18) england(44.77) saint(40.23) america(40.14) holy(35.98) pope(32.87) priest(29.76) russian(29.75) culture(28.43) christianity(27.87) religious(27.10) reformation(25.39) ukrainian(23.20) mary(22.86) belong(21.83) bishop(21.57) anglican(18.19) rite(18.16) teaching(16.50) christian(15.57) diocese(15.44) ...

2. sense: church, church_building "a place for public (especially Christian) worship; "

house(1733.29) worship(1079.19) building(620.77) mosque(529.07) place(507.32) synagogue(428.20) god(408.52) kirk(368.82) build(93.17) construction(47.62) street(47.18) nation(41.16) road(40.12) congregation(39.74) muslim(37.17) list(34.19) construct(31.74) welcome(29.23) new(28.94) prayer(24.48) temple(24.40) design(24.25) brick(24.24) erect(23.85) door(20.07) heaven(19.72) plan(18.26) call(17.99) renovation(17.78) mile(17.63) gate(17.09) architect(16.86) conservative(16.46) situate(16.46) site(16.37) demolition(16.16) quaker(15.99) fort(14.59) arson(12.93) sultan(12.93) community(12.88) hill(12.62) ...

3. sense: church_service, church "a service conducted in a church; "

service(5225.65) chapel(1058.77) divine(718.75) prayer(543.96) hold(288.08) cemetery(284.48) meeting(271.04) funeral(266.05) sunday(256.46) morning(169.38) attend(143.64) pm(133.56) meet(115.86) conduct(98.96) wednesday(90.13) religious(89.19) evening(75.01) day(74.45) friday(73.17) eve(70.01) monday(67.96) cremation(64.73) saturday(60.46) thursday(60.46) june(57.78) tuesday(56.08) crematorium(55.53) weekly(53.36) procession(50.53) burial(48.60) december(48.46) ceremony(46.47) september(46.10) interment(42.31) lead(38.79) family(34.19) deceased(31.73) visitation(31.44) ...

Figure 1: Fragment of the topic signatures for the three senses of church. The values in parenthesis correspond to the strength of the relevance. Only the top scoring terms are shown.

2.1. Definition of monosemy

We say that a word is monosemous if it has a unique sense, that is, if a word has a unique *synset* taking into account all its parts of speech. Following our definition, a word that has a unique sense as a noun does not need to be monosemous: it is monosemous only if it does not have any sense in the other parts of speech. For instance, the word *cure* has a single sense as a noun in WordNet (version 1.6), but it also has two senses as a verb. Therefore, we consider *cure* a polysemous word.

2.2. Method followed to collect examples

In this work we have followed the monosemous relatives method, as proposed in (Leacock et al., 1998). This method uses monosemous synonyms or hyponyms to construct the queries. For instance, the first sense of *channel* in Figure 1 has a monosemous synonym "*transmission channel*". All the occurrences of "*transmission channel*" in any corpus can be taken to refer to the first sense of channel. In our case we have used the following kind of relations in order to get the monosemous relatives: hypernyms, direct and indirect hyponyms, and siblings. The advantages of this method is that it is simple, it does not need error-prone analysis of the glosses and it can be used with languages where glosses are not available in their respective WordNets.

Google[3] was used to retrieve the occurrences of the monosemous relatives. In order to avoid retrieving full documents (which is time consuming) we take the context from the snippets returned by Google. (Agirre et al., 2001) showed that topic signatures built from sentence context were more precise than those built from full document context, provided the amount of sentence contexts was larger.

The snippets returned by Google (up to 1,000 per query) are processed, and we try to extract sentences (or fragments of sentences) containing the search term from the snippets. The sentence (or fragment) is usually marked by three dots in the snippet. Some of the potential sentences are discarded, according to the following heuristics: length shorter than 6 words, the number of non-alphanumeric characters is greater than the number of words divided by two, or the number of words in uppercase is greater than those in lowercase.

Table 1 shows some figures for the snippets and filtered examples. The snippets columns show the amount of monosemous and polysemos words in WordNet and the size and number of retrieved examples. The other two columns show the number of examples after detecting the sentence and discarding some of the examples.

3. Method to build the topic signatures

A topic signature is a vector, as shown in Equation 1, where t is the topic (i.e the target word sense) and w_i is a related word with its relatedness weight s_i.

$$\{t, < (w_1, s_1), (w_2, s_2) \dots (w_i, s_i) \dots >\} \qquad (1)$$

As explained in previous sections we can build these vectors from a sense-tagged corpora, observing which words co-occur distinctively with each sense, or we can try to aqcuire examples automaticaly (i.e web) with the monosemous relatives method and associate these acquired documents to each target word senses.

The method to construct topic signatures proceeds as follows: (a) We first organize the examples collected from

[3]We use the offline XML interface kindly provided by Google.

	snippets		filtered examples	
	monosemous	polysemous	monosemous	polysemous
number of words	91,884	15,607	91,884	-
number of examples	62,745,798	13,869,675	14,664,798	-
size in word	1,678,759,964	376,335,658	307,332,541	-
average examples per word	682.87	888.68	159.60	-
average size (words) per word	18,270.42	24,110.17	3,344.78	-
average size (words) per example	26.75	27.13	20.95	-

Table 1: Statistics for the examples retrieved from the Web

number of polysemous words	15,875
number of senses	37,678
average senses per word	2.38

	1. meth	2. meth	3. meth	4. meth	Total
number of examples	2,728,082	11,145,888	12,028,151	109,138,720	135,040,841
average examples per sense	72.4	295.8	319.2	2,896.6	3,584.1
average examples per word	172.3	703.9	759.6	6,892.2	8,527.9

Table 2: Data for the examples gathered for the senses of polysemous words using the mosemous relatives method.

the web in collections, one collection per word sense. (b) For each collection we extract the words and their frequencies, and compare them with the data in the collections pertaining to the other word senses using the $tf.idf$ statistic. (c) The words that have a distinctive frequency for one of the collections are collected in a list, which constitutes the topic signature for the respective word sense.

Optionally: (d) The topic signatures for the word senses are filtered with the cooccurrence list of the target word taken from balanced corpora such as the BNC. This last step takes out some rare and low frequency words from the topic signatures.

3.1. One collection per word sense

For each sense of a polysemous noun, we gather all examples of its monosemous relatives, including synonyms, hypernyms, siblings and hyponyms (including indirect hyponyms). The intuition is that relatedness decreases with the distance, so we assigned a numeric value to each of them: synonyms are assigned 1, hypernyms 2 and siblings 3. Hyponyms get a value according to the distance: direct hyponyms are assigned 1, second level hyponyms 2, etc. The maximum weitght is 4.

Table 2 shows the amount of examples gathered for each sense of the polysemous words, listed according to the method used. On average we gather 8,526 examples per polysemous word, and each of its senses gets 3,584 examples.

3.2. Weighting the words in context

In the previous step we constructed vectors of frequencies. Frequencies are not good indicators of relevancy, so different functions can be used in order to measure the relevance of each term appearing in the vector corresponding to one sense in contrast to the others. That is, terms occurring frequently with one sense, but not with the other senses of the target word, are assigned high weights for the associated word senses, and low values for the rest of word

		s. 1	s. 2	s. 3
num. of examples	1.met	0	330	112
	2.met	727	274	0
	3.met	2,916	1,203	787
	4.met	1,801	2,060	870
	Total	5,444	3,867	1,769
signature size (words)		9,079	7,757	4,450

Table 3: Statistics for the three senses of noun *church*

total size	228,331,038
total size (non zero weights)	158,099,870
average size per signature	6,623.5
average size (non zero weights)	4,587.2

Table 4: Size in word of the topic signatures

senses. Terms occurring evenly among all word senses are also assigned low weights for all the word senses. In this work we use $tf.idf$ (see 2) which yielded the best results in (Agirre and Lopez, 2003).

$$tf.idf = \frac{tf_t}{max_t tf_t} \times \log \frac{N}{df_t} \quad (2)$$

The topic signatures are constructed assigning these weights to the words in the context of each of the word sense. Figure 1 shows the topic signatures for the three senses of church, and 3 shows the number of examples according to the numeric value of the relative for each of the three senses, alongside the size of each of the signatures. Table 4 shows the total and average sizes of the acquired signatures. We couls also filter out the most irrelevant words from the signature, and the same table shows the figures if the words with zero $tf.idf$ are removed.

3.3. Filtering

Hand inspection of the automatically constructed topic signatures show that some weird words get high weights.

The snippets gathered from the Web, and the fact that we compare each word sense against the others can produce high weights for some rare terms.

This effect can be reduced in the following way: we collect contexts of occurrences for the target *word* from a large and balanced corpus, and select the words that are highly related to the word. This list of words related to the target word is used in order to filter all topic signatures corresponding to the target word, that is, context terms which are not relevant for the target word are deleted from the topic signature.

Our experience with topic signatures shows that filtering makes the topic signatures more pleasent to the eye, but it does not have much effect in performance. For instance, (Agirre et al., 2004) shows similar correlation values for topic signatures constructed with and without filtering when comparing simmilarity methods. Due to the computational effort needed to do the cleaning process, we constructed the topic signatures withour filtering.

4. Conclusions

This paper reports the construction of a publicly available resource which includes for each nominal word sense en Wordnet 1.6 both automatically extracted examples and topic signatures built based on those examples. The size of the topic signatures thus constructed is of around 4,500 words per word sense. The topic signatures used in this paper can be browsed in full[4], and are publicly available for download[5].

For the future, we plan to release topic signatures for newer releases of WordNet.

5. Acknowledgments

This research has been partially funded by the Spanish Research Department (HERMES TIC2000-0335-C03-02) and by the European Commission (MEANINGIST-2001-34460).

6. References

Agirre, E., E. Alfonseca, and O. Lopez, 2004. Approximating hierachy-based similarity for wordnet nominal synsets using topic signatures. *Proc. of the 2nd Global WordNet Conference*.

Agirre, E., O. Ansa, E. Hovy, and D. Martinez, 2000. Enriching very large ontologies using the www. *Proceedings of the Ontology Learning Workshop, ECAI*.

Agirre, E., O. Ansa, D. Martinez, and E. Hovy, 2001. Enriching wordnet concepts with topic signatures. *Proceedings of the NAACL worshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*.

Agirre, E. and O. Lopez, 2003. Clustering wordnet word senses. *Proceedings of the Conference on Recent Advances on Natural Language (RANLP'03)*.

Agirre, Eneko and David Martinez, 2004. The effect of bias on an automatically-built word sense corpus. *Proceedings of the 4rd International Conference on Languages Resources and Evaluations (LREC)*.

Alfonseca, E. and S. Manandhar, 2002. Extending a lexical ontology by a combination of distributional semantics signatures. *Lecture Notes in Computer Science*, 2473.

Atserias, Jordi, Luís Villarejo, German Rigau, Eneko Agirre, John Carroll, Bernardo Magnini, and Piek Vossen, 2004. The meaning multilingual central repository. In *Second International WordNet Conference-GWC 2004*. Brno, Czech Republic. ISBN 80-210-3302-9.

Leacock, Claudia, Martin Chodorow, and George A. Miller, 1998. Using corpus statistics and wordnet relations for sense identification. *Computational Linguistics*, 24(1):147–165.

Lin, C. and E. Hovy, 2000. The automated acquisition of topic signatures for text summarization. *Proceedingd of the COLING Conference*.

Maedche, A. and S. Staab, Forthcoming. Ontology learning. *Handbook of Ontologies in Information Systems, editors S. staab and R. Studer*.

Mihalcea, Rada, 2002. Bootstrapping large sense tagged corpora. *Proceedings of the 3rd International Conference on Languages Resources and Evaluations (LREC)*.

Miller, George, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller, 1990. Five papers on WordNet. CSL Report 43, Cognitive Science Laboratory, Princeton University.

[4] `http://ixa3.si.ehu.es/cgi-bin/signatureak/signaturecgi.cgi`

[5] `http://ixa2.si.ehu.es/pub/webcorpus`