The web as a resource for WSD[1]

Julio Gonzalo **Irina Chugur** **Felisa Verdejo**

Departamento de Lenguajes y Sistemas Informáticos
UNED (Madrid-Spain)
{julio,irina,felisa}@lsi.uned.es

Abstract

We review main approaches to acquire lexical information using the web, focusing on the automatic acquisition of sense-tagged corpora, which could turn out to be the most promising way of solving the knowledge-acquisition bottleneck of supervised Word Sense Disambiguation systems.

1 Introduction

Why general Word Sense Disambiguation (WSD) systems are not (to the best of our knowledge) being used in real-world applications? In our opinion, some plausible reasons are:

- General Word Sense Disambiguation, as an intermediate task, is frequently harder than the final application. This is probably the case of the most commonly mentioned WSD applications: mono and multilingual Information Retrieval and Machine Translation.
- Different applications demand different sense distinctions and different sense granularities.
- Unsupervised systems still have poor performance...
- ... and supervised systems, in general, hardly have resources to be supervised with. Currently, supervised systems can only attempt an “all-words” WSD task in English, and with very scarce resources.

In order to find out whether WSD systems can play a significant role in Human Language Technologies, a necessary step is to make supervised WSD algorithms applicable, finding ways of acquiring lexical information and, in particular, building training corpora at a low cost.

The size, heterogeneity and multilingual character of the web, combined with the coverage and efficiency of web search engines, are a natural path to explore automatic ways of acquiring such information. In this paper we review some of the approaches attempted over the last few years. In Section 2, we review strategies to mine lexical information from the web that can be used for WSD. In Section 3, we focus on the direct acquisition of training samples from the web. Finally, in Section 4 we draw some conclusions.

2 Mining lexical information from the web

2.1 Domain information

Wordnet, which is the most frequently used sense inventory in WSD, does not incorporate topical information, which is very valuable for sense disambiguation and for many other purposes. We will mention here two strategies to enrich Wordnet with domain information: extraction of topical signa-

[1] The contents of this presentation are essentially included as a part of “Resources for Word Sense Disambiguation”, a chapter of the book: Agirre and Edmonds (eds.) *Word Sense Disambiguation: Algorithms, applications and trends*, to be published by Kluwer AP.

tures, and association of Wordnet senses with web directories.

Topic signatures

In (Agirre et al. 00), the web is used to enrich wordnet senses with topical signatures. A topic signature is defined as a list of words which are topically related to the word sense, together with a measure of the strength of the association. An example of the authors is "waiter" as a person who waits vs. as a person who serves a table. In the first sense, the topic signature could be made of words such as *hospital, station, airport, cigarette*, etc. In the second sense, the list would rather include words such as *restaurant, menu, waitress, dinner*, etc.

Such topic signatures are built in two main steps: 1) using Altavista to retrieve sets of documents associated each word sense, and 2) using the documents to extract and weight the words that form the topic signatures for every sense:

In step one, a list of cuewords for each sense is extracted from wordnet (including synonyms, words in the gloss and words in related synsets). Then, for each sense a boolean query is formed to retrieve documents containing the original word, at least one of the cuewords of the intended sense, and none of the cuewords for the other senses of the word.

Then, in the second step, each word in the set of documents related to one word sense is assigned a weight that grows when the frequency of the word is higher than what would be expected from the contrast set made of the documents belonging to the other senses of the word. The words and their weights, in decreasing order of weight, form the topic signature for each word sense.

In this work, the topic signatures are used in a straightforward WSD approach (to test the utility of the information provided by the signatures) with encouraging results. They are also used to cluster wordnet senses (two close senses will have close topic signatures), which are in turn successfully incorporated to the WSD strategy based solely on topic signatures. The authors conclude that the quantitative evidence in favor of topic signatures is high, but a qualitative inspection of the data suggests that more filtering is needed to discard poor quality documents and some topical biases of the web (e.g. the topic signature for boy is too closely related to pornography issues). Unfortunately there has not yet been a large-scale application of these techniques to enrich the full wordnet (rather than a sample for evaluation purposes).

Association of web directories to word senses

In (Santamaría et al., 2003), Wordnet senses are automatically associated to web directories. Web directories, (such as Yahoo or ODP) are hierarchical thematic categories that organize the information in the web so that the information of interest to a user can be located not only *querying* (as in a search engine), but also *browsing* the contents of the web by iterative topic refinement. The most interesting feature of web directories, from the perspective of the web as a corpus, is that both the directories and the association of web pages to directories are manually constructed. Compared to the full web, then, directories should be a much cleaner and balanced source of information. The hypothesis of Santamaría is that one or more assignments of web directories to a word sense would be an enormously rich and compact source of topical information about the word sense, which includes both the hierarchy of associated subdirectories and the web pages beneath them.

The approach consists mainly of three stages:

First, a query is formed similarly to (Agirre et al. 00), using relevant cuewords extracted from Wordnet for every word sense, and using cuewords from the other senses as negative information.

The query is launched against ODP (www.odp.com) directories, and a set of directories (rather than documents) is retrieved.

Then the directories are compared with the word senses, assuming that a relevant directory (represented by the chain of parent directories that lead to it) will have some degree of overlapping with the word sense (represented by the chain of hypernyms of the associated synset in Wordnet). The authors apply a set of additional criteria and filters to end up with possible associations and an empirical confidence measure for each association.

The result of the algorithm is not only a set of (word sense, ODP directory) associations, but also 1) directories classified as hyponyms of a given sense, e.g. "integrated circuit" as a hyponym of "circuit" in the "electric circuit" sense; 2) potential sense clusters (from a topical point of view), when

a single directory is assigned to two or more senses of a word; and 3) discarded directories, which can be subsequently mined to discover potentially relevant senses which are not included in Wordnet but can be relevant in some domains (e.g. Java as a programming language, Tiger as the golfist or Oasis as the music band).
A direct application of the algorithm to all (single term) nouns in Wordnet gives highly accurate associations with a relatively low coverage (partly because not every noun in wordnet has domain specificity). But these associations can be inherited by words in the same synset, and possibly also by words in hyponym synsets, which should substantially increase the coverage. The results of this work can be downloaded in http://nlp.uned.es/ODP. The usefulness of sense/directory association has been tested in a sense-tagged corpora acquisition task, which is commented in Section 3.3.

2.2 Parallel Corpora

Sometimes choosing a correct translation for a word in context can be easier than disambiguating its sense. This is often the case when there are enough translation statistics extracted from available parallel corpora. In such cases, translation information can be used to partially disambiguate the word, because only a subset of the possible senses can be translated with a given term (Gale, Church et al. 1992). Applying this criterion on several languages, the combined translation restrictions should fully disambiguate the word; otherwise, the remaining sense candidates are so close that no language lexicalizes the difference, hence probably they do not need to be distinguished for any practical application (Resnik & Yarowsky 99).
The problem with this approach is again the knowledge acquisition bottleneck: parallel corpora are scarce resources, specially when English is not one of the languages involved. And again, the web can be mined for parallel corpora, thus enabling Machine Translation technologies based on statistical translation for a much broader set of languages and domains.

Creating a parallel corpus out of the web usually involves three steps (Resnik & Smith 2002): a) locating domains, sites or pages that might have parallel translations; b) generation of candidate pairs from such data; and c) filtering candidate pairs with structural or content-based criteria.

Generation of candidate pairs can be done with relatively simple strategies: language identification, URL matching (e.g. substituting "esp" – spanish - with "eng" –english - in an existing URL and checking out whether the substituted URL also exists), comparison of document lengths, etc.

Filtering candidate pairs can be done mainly according to structural criteria (is the structure of both documents similar?) or content criteria (do they have a similar content?). Relevant approaches include:
PTMINER (Chen & Nie 2000) locates promising sites by querying for pages in a given language that contain links to pages in different languages. Once bilingual sites are located and crawled, filtering criteria include language identification, URL matching and length comparison, without structural or content comparison. An English-French corpus of around 100Mb per language, produced with these techniques, has been succesfully used to improve Cross-Language Information Retrieval (CLIR) systems by participants in the CLEF comparative evaluation of Multilingual Information Retrieval.
BITS (Ma and Liberman 1999) use bilingual dictionaries to compute a content-based similarity score between candidate pairs, with additional filters for document length, similarity of anchors (numbers, acronyms, etc), etc.
STRAND (Resnik 1999) uses structural filtering to compare language pairs, linearizing the HTML structure of both documents and aligning the resulting sequences. Four scalar values on the alignment characterize the quality of the alignment, and a Machine Learning process is used to optimize filtering according to these parameters, to obtain a precision of .97 and a recall of .83 over a set of English-French candidate pairs. In (Resnik & Smith 2002), STRAND is enhanced with content-based similarity measures and applied over the Internet Archive (www.archive.org) to obtain an English-Arabic parallel corpus of more than 1M tokens per language, with a precision of .95 and a recall of .99 over the extracted candidate pairs. An interesting feature of STRAND, when combined with the Internet Archive, is that it solves legal distribution problems by listing the URLs rather

than the documents themselves, and that URLs are stable as part of the Internet Archive. These issues are not commonly addressed in the literature of web mining for language resources.

Besides parallel corpora, the evidence about translation in context can also be obtained from comparable corpora or even from the web as a big, comprehensive multilingual corpus. (Grefenstette 1999) showed that multiword translation can be done accurately using the co-occurrence statistics of the candidate translation pairs for the original words in the multiword expression. For instance, "strong tea" is much more frequent in the web than "powerful tea" according to the statistics of querying Altavista. This principle has been fully exploited to align Spanish and English noun phrases using evidence from the CLEF document collection (Peters et al. 2002). The large scale bilingual alignment (over 1M different phrases in each language) has been succesfully applied to obtain indicative cross-language pseudo summaries (López-Ostenero et al 2002) and to assist query formulation and refinement (López-Ostenero et al. 2003) in interactive Cross-Language Information Retrieval experiments. The cross-language phrase mapping could also be used as evidence for WSD, and an extension to web extracted comparable corpora could be very useful to reach domain independence and a larger set of usable languages. Even without using the web, the CLEF test suite (Peters et al. 2002) already comprises a few gigabytes of comparable corpora in eight European languages (English, Spanish, German, Italian, French, Finnish, Swedish and Dutch): this is an enormously rich resource that has not yet been used for WSD purposes.

3 Automatic acquisition of sense-tagged corpora

The most direct way of using the web to enhance WSD performance is the automatic acquisition of sense-tagged corpora from the web, as the fundamental resource to train supervised WSD algorithms. Although this kind of strategy is far from being commonplace in the WSD literature, there has already been a number of different and potentially useful strategies to achieve this goal, which we review here.

3.1 Acquisition by direct web searching.

In (Leacock et al. 1998), the monosemous lexical relatives of a word sense provide a key for finding training sentences in a corpus. For instance, looking for "business suit" as a monosemous hyponym of "suit" can give us training sentences for the adequate sense of suit. In (Mihalcea and Moldovan 99) this idea is extended to a) query the full web and b) using other useful cuewords in WordNet.

Mihalcea and Moldovan use four ranked procedures to search the web for instances of a word sense: first, monosemous synonyms are tried; then, defining phrases; in case of failure, a boolean query is made with synonyms (grouped with AND operators) and words from the defining phrases (using the NEAR operator). Finally, synonyms and words from the defining phrases are grouped with the AND operator. The method is shown to be highly productive (an average of 670 examples per word sense in the sample chosen for evaluation) and precise (91% of the examples acquired were correct).

With such good results, why the use of the web to extract sense-tagged corpora did not immediately became a mainstream approach? In (Agirre and Martínez 2000), the authors replicated the same strategy to build a sense-tagged corpora and used the results to train a WSD system that was tested against a subset of Semcor. The results were disappointing: only a few words get better than random results. Agirre and Martínez concluded that the examples, being themselves correct, could provide sistematically misleading features, and that the unbalanced number of examples (all word senses have basically the same number of training instances) could also mislead the algorithm. In our opinion, another problem of direct querying of the web to get training samples is that we will only capture a fraction of the relevant examples, the ones that co-occur with terms related to the word sense via WordNet relations. This set of examples may not be even a significant fraction of the uses of the word sense.

3.2 Bootstrapping.

In (Mihalcea 2002), the method described in the previous section is enriched with a bootstrapping approach inspired in (Yarowsky 95), where a few tagged samples are used to train a decision list, which is then employed to tag new instances. In this paper, Mihalcea creates a set of seeds extracted from Semcor, WordNet and a more restricted version of (Mihalcea and Moldovan 1999). Then, the web is searched using queries formed with the seed expressions. Finally, the words surrounding the seed expressions are disambiguated using the algorithm in (Mihalcea and Moldovan, 2000), which in turn serve as new seed expressions for a new bootstrapping iteration. The sense-tagged corpus generated with this approach was tested in the Senseval 2 WSD task, with excellent results: the system performed the best both in the English *lexical sample* and *all words* tasks, and a good part of the success is due to the web acquired corpora. For instance, in the all-words task, the first sense heuristic gives 63.9% precision; if only Semcor and WordNet are used for training, the result is 65.1% (+ 1.2 absolute improvement). The same algorithm, trained with the web-based corpus, achieves 69.3% precision (+ 5.4 absolute improvement).

3.3 Acquisition via web directories.

The (word sense, web directory) associations obtained in (Santamaría et al. 2003) can be trivially applied to obtain sense-tagged corpora, extracting the occurrences of the word in the web pages associated to the web directory or in the manually built description of the pages under the directory. Comparing to the strategies described above, the use of directories has, a priori, at least three advantages: 1) catalogued web pages are a cleaner source of information than the full web; 2) as the algorithm retrieves directories rather than documents, the occurrences of the word in the documents associated to the directory do not necessarily co-occur with the seed words used in the web search, permitting a larger variety of training samples; 3) web directories can be distributed without copyright problems, and they are more stable in time than individual web pages. The counterpart of the method is that it can only be applied to word senses which can be related to topical domains, which is not the case for every word sense in WordNet.

In (Santamaría et al. 2003), a preliminary experiment is conducted using the nouns in the Senseval 2 test suite, and only the examples found in the pages that describe the web categories (rather than following the links to the web pages listed under the category). The experiment showed that when the number of training instances is similar, the examples acquired automatically work as well as the manual examples provided in Senseval 2 for training purposes. Again the problem is coverage: with this restrictive approach, the overall number of training instances is substantially lower than the original Senseval 2 training material.

3.4 Web-based cooperative annotation.

Finally, an alternative that uses the web but does not acquire sense-tagged corpora automatically is the *Open Mind Word Expert*, in which a web site collects sense annotations made by web users (Chklovski and Mihalcea 2002). The system has an active learning component, which uses current annotations to train a WSD system, and selects the harder examples as the next examples to be offered to the users of the system. The system has some "game-like" features to engage users, including a "Hall of Fame" for the most active contributors. At the time of writing this review, the Open Mind Word Expert has already collected over 70,000 human-annotated instances, which is roughly the same amount of Semcor instances for polysemous nouns.

4 Conclusion

The work that we have briefly reviewed here indicates, in our opinion, that the web can succesfully be used as a source of information to acquire semantic information, in general, and training examples in particular. We believe that much more attention should be given to this topic in the near future, as one of the primary ways of scaling WSD technologies to fit application needs.

References

Agirre, E., Ansa, O., Hovy, E. and Martínez, D. Enriching very large ontologies using the WWW. In Proc. of the Ontology Learning Workshop, ECAI, Berlin, Germany, 2000.

Agirre E. and D. Martínez (2000). Exploring automatic word sense disambiguation with decision lists and the Web. *Proceedings of the Semantic Annotation And Intelligent Annotation workshop organized by COLING* Luxembourg.

Agirre E. and D. Martínez (2001a). Decision Lists for English and Basque. Proceedings of the

SENSEVAL-2 Workshop. In conjunction with ACL'2001/EACL'2001. Toulouse, France.

Chen, J. and Nie, J. (2000). Web parallel text mining for chinese english cross-language information retrieval. In *Proc. of the International Conference on Chinese Language Computing*.

Chklovski, T. and Mihalcea, R. (2002). Building a Sense Tagged Corpus with Open Mind Word Expert. In Proceedings of the ACL'02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions.

Gale, W., Church, K. and Yarowsky, D. (1992) One Sense per Discourse, Proc. of the DARPA Speech and Natural Language Workshop.

Grefenstette, G. (1999). The WWW as a Resource for Example-Based MT Tasks, ASLIB'99 Translating and the Computer 21.

Kilgarriff, A. and M. Palmer, guest editors (2000). Special Issue on Senseval. Computers and the Humanities (34(1-2)).

Kilgarriff, A. (2001). SENSEVAL-2, Toulouse, France.

Kilgarriff, A. (2002) English Lexical Sample Task Description. Proceedings of Senseval-2 ACL Workshop.

Kilgarriff, A. and Grefenstette, G. (eds.) (2003) *Computational Linguistics, Special Issue on the Web as Corpus*, to appear.

Leacock, C., Chodorow, M. and Miller, G. (1998) Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics.*

López-Ostenero, F., Gonzalo, J., Peñas, A. and Verdejo, F. (2002) Noun Phrase Translations for Cross-Language Document Selection. In *Cross-Language Information Retrieval and Evaluation: Revised papers of the CLEF 2001 workshop*, Springer-Verlag LNCS Series.

López-Ostenero, F., Gonzalo, J., Peñas, A. and Verdejo, F. (2003). Phrases are better than words for Interactive Cross-Language Query formulation and Refinement. In *Cross-Language Information Retrieval and Evaluation*: *Revised papers of the CLEF 2002 workshop*, Springer-Verlag LNCS Series, to appear.

Ma, X. and Liberman, M. (1999). BITS: a method for bilingual text search over the web. In *Proc. of the Machine Translation Summit VII.*

Mihalcea, R. and D. Moldovan (1999). An Automatic Method for Generating Sense Tagged Corpora. In *Proc. AAAI'99.*

Mihalcea, R. and Moldovan, D. (2000) An Iterative Approach to Word Sense Disambiguation. In *Proceedings of Flairs 2000*.

Mihalcea, R. (2002) Bootstrapping large sense-tagged corpora. In *Proc. LREC'02*.

Peters, C., Braschler, M., Gonzalo, J. and Kluck, M. (eds) (2002). *Cross-Language Information Retrieval and Evaluation. Revised papers of the CLEF 2001 workshop*. Springer-Verlag LNCS Series.

Resnik, P. (1999). Mining the web for bilingual text. In *Proceedings of ACL'99*.

Resnik, P. and Yarowsky, D. (1999). "Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation", Natural Language Engineering 5(2), pp. 113-133.

Resnik, P. and Smith, N. (2002). The Web as a Parallel Corpus, University of Maryland technical report UMIACS-TR-2002.

Santamaría, C., Gonzalo, J. and Verdejo, F. (2003) Automatic association of web directories to word senses. *Computational Linguistics, Special Issue on the Web as Corpus*, to appear.

Sussna, M (1993). Word Sense Disambiguation for Free Text Indexing Using a Massive Semantic Network. In Proceedings of the Second International Conference on Information and Knowledge Base Management, CIKM'93.

Yarowsky, D. (1995). Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In Proceedings of ACL'95.