# The effect of bias on an automatically-built word sense corpus

**David Martinez, Eneko Agirre**

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

**Abstract**

The goal of this paper is to explore the large-scale automatic acquisition of sense-tagged examples to be used for Word Sense Disambiguation (WSD). We have applied the "monosemous relatives" method on the Web in order to build such a resource for all nouns in WordNet. The analysis of some parameters revealed that the distribution of the word senses (bias) in the training and test corpus is a determinant factor. Provided there is a method to approximate the bias for each word sense, the results we obtained for English are comparable to the use of hand-tagged data (Semcor), which is a very interesting perspective for lesser studied languages.

## 1. Introduction

The results of recent WSD exercises, e.g. Senseval-2[1] (Edmonds and Cotton, 2001) show clearly that Word Sense Disambiguation (WSD) methods based on hand-tagged examples are the ones performing best. However, one of the main drawbacks for supervised WSD is the acquisition bottleneck, as the systems need large amounts of costly hand-tagged data. The situation is more dramatic for lesses studied languages. In order to overcome this, different research lines are being pursued: automatic acquisition of training examples, bootstrapping techniques (Yarowsky, 1995), or active learning (Argamon-Engelson and Dagan, 1999). In this work, we have focused on the automatic acquisition of examples.

When supervised systems have not specifically worked training examples for a target word, they need to rely on publicly available all-words sense-tagged corpora, like Semcor (Miller et al., 1993), which is tagged with WordNet word senses. The best performing systems that participated in the English all-words task in Senseval-2 were supervised systems trained on Semcor. Unfortunately, for many words, this corpus has only a handful of tagged examples . In fact, only a few systems could overcome the Most Frequent Sense baseline, which would tag each word with the sense occurring most frequently in Semcor. For our approach, we will also use Semcor as a resource, both for training examples and as an indicator of the distribution of the senses of the target word.

The goal of our experiment is to evaluate up to which point we can automatically acquire examples for word senses and train accurate supervised WSD systems on them. The method we applied is based on the monosemous relatives of the target words (Leacock et al., 1998), and we studied some parameters that affect the quality of the acquired corpus: the distribution of the number of training instances per each word sense (bias), the substitution or not of the monosemous relative for the target word, and the type of features used for disambiguation (local vs. topical).

In (Leacock et al., 1998), the method to obtain sense-tagged examples using monosemous relatives is presented. In this work, they retrieve the same number of examples per sense, using local content words and topical features, and they give preference to monosemous relatives that consist in a multiword containing the target word. Their experiment is evaluated on 3 words (a noun, a verb, and an adjective) with coarse sense-granularity and few senses. The results showed that the monosemous corpus could provide precision comparable to hand-tagged data.

In another related work, (Mihalcea, 2002) generated a sense tagged corpus (GenCor) by using a set of seeds consisting of sense-tagged examples from four sources: SemCor, WordNet, examples created using the method above, and hand-tagged examples from other sources (e.g., the Senseval-2 corpus). By means of an iterative process, the system obtained new seeds from the retrieved examples. She reported a clear improvement of performance in the Senseval-2 all-words task using the automatically acquired corpus. An experiment in the lexical-sample task showed that the method was useful for some words.

This paper is structured as follows. Section 2 introduces the experimental setting for evaluating the acquired corpus. Section 3 is devoted to the process of building the corpus, which is evaluated in Section 4. Finally, some conclusions are given in Section 5.

## 2. Experimental Setting for Evaluation

In this section we will present the Machine Learning method, the features used to represent the context, the two hand-tagged corpora used in the experiment and the word-set used for evaluation.

### 2.1. Decision Lists

The learning method used to measure the quality of the corpus is **Decision Lists** (DL). This algorithm is described in (Yarowsky, 1995). In this method, the sense with the highest weighted feature is selected, according to its log-likelihood (see Formula 1). The cases where the denominator is zero are smoothed by the constant 0.1 .

$$\arg\max_k w(s_k, f_i) = \log(\frac{Pr(s_k|f_i)}{\sum_{j\neq k} Pr(s_j|f_i)}) \qquad (1)$$

### 2.2. Features

In order to represent the context, we used a set of features frequently used in the literature for WSD tasks (Agirre and Martinez, 2000). We distinguish two types of features:

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

- Local features: Bigrams and trigrams, formed by the word-form, lemma, and part-of-speech of the surrounding words. Also the content lemmas in a ±4 word window around the target.
- Topical features: All the content lemmas in the context.

We have analyzed the results using local and topical features separately, and also the combination of both types.

### 2.3. Hand-tagged corpora

The performance of the WSD system when trained on the automatic sense-tagged corpus was compared with that of the same system trained on Semcor.

For evaluation, the testing part of the English lexical-sample task was chosen. The advantage of this corpus was that we could focus in a word-set with enough examples for testing. Besides, it is a different corpus, so the evaluation is more realistic than that made using cross-validation. The testing examples whose senses were multiwords or phrasal verbs were removed.

It is important to note that the training part of Senseval-2 lexical-sample was not used in the process, as our goal was to test the performance we could achieve with the minimal resources (i.e. those available for any word).

### 2.4. Word-set

The experiments were performed on the 29 nouns available for the Senseval-2 lexical-sample task. We will separate these nouns in 2 sets, depending of the number of examples they have for training: Set A will contain the 16 nouns with more than 10 examples in Semcor, and Set B the remaining low-frequency words.

## 3. Building the monosemous relatives web corpus

In order to build this corpus[2], we have acquired 1000 Google snippets for each monosemous word in WordNet 1.7. Then, for each word sense of the ambiguous words, we gathered the examples of its monosemous relatives. This method is inspired in (Leacock et al., 1998), and has shown to be effective in experiments of topic signature acquisition (Agirre and Lopez, 2004). This last paper also shows that it is possible to gather monosemous relatives for all noun senses in WordNet[3].

The basic assumption is that for a given word sense of the target word, if we had a monosemous synonym of the word sense, then the examples of the synonym should be very similar to the target word sense, and could therefore be used to train a classifier of the target word sense. The same, but in a lesser extent, can be applied other monosemous relatives, such as direct hyponyms, direct hypernyms, siblings, indirect hyponyms, etc. The expected reliability decreases with the distance in the hierarchy from the monosemous relative to the target word.

The monosemous-corpus was built using the simplest technique: we collect examples from the web for each of the monosemous relatives. The relatives have an associated number (distance), which indicates their relevance: the higher the distance, the less reliable the relative. A sample of monosemous relatives for some senses of *church* is shown below:

Synonyms (0): *church building*
Direct hyponyms (1): *Protestant Church*
Direct hypernyms (2): *house of prayer*
Distant hyponyms (3,4,...): *Western Church*
Siblings (3): *Hebraism*

### 3.1. Collecting the examples

The examples are collected following these steps

**1:** We query Google[4] with the monosemous relatives for each sense, and we extract the snippets as returned by the search engine. All snippets returned by google are used (up to 1000). The list of snippets is sorted in inverse order to the results as retrieved by the search engine. This is done because the top hits usually are titles and incomplete sentences that are not very useful.

**2:** We extract the sentences (or fragments of sentences) around the target search term. Some of the sentences are discarded, according to the following criteria: shorter than 6 words, having more non-alphanumeric characters than words/2, or having more words in uppercase than in lowercase.

**3:** The automatically acquired examples contain a monosemous relative of the target word. In order to use these examples to train the classifiers, the monosemous relative (which can be a multiword term) is substituted by the target word. In the case of the monosemous relative being a multiword that contains the target word (e.g. *Protestant Church*) we can choose not to substitute, because *Protestant* can be a useful feature for the first sense of church. In these cases, we decided not to substitute and keep the original sentence, as our preliminary experiments suggested.

**4:** For a given word sense, we collect the desired number of examples (see following section) in order of type: we first collect all examples of type 0, then type 1, etc. up to type 3 until the necessary examples are collected. We did not collect examples from type 4 upwards. We did not make any distinctions between the relatives from each type. (Leacock et al., 1998)) give preference to multiword relatives containing the target word, which could be a better approach for future work.

On average, we have acquired roughly 40.000 examples for each of the target words used in this experiment.

### 3.2. Number of examples per sense (bias)

Previous work (Agirre and Martinez, 2000) has reported that the distribution of the number of examples per word sense (bias for short) has a strong influence in the quality of the results. That is, the results degrade significantly

[2] The automatically acquired corpus will be referred indistinctly as web-corpus, or monosemous-corpus

[3] All the examples in this work are publicly available in `http://ixa2.si.ehu.es/pub/webcorpus`

[4] We use the offline XML interface kindly provided by Google for research

| Sense | Semcor | | Web bias | | Proportional | | Minimum ratio | | Senseval testing | |
|---|---|---|---|---|---|---|---|---|---|---|
| | # ex | % | # ex | % | # ex | % | # ex | % | # ex | % |
| authority#1 | 18 | 60 | 338 | 0,5 | 338 | 33,7 | 324 | 59,9 | 37 | 40,7 |
| authority#2 | 5 | 16,7 | 44932 | 66,4 | 277 | 27,6 | 90 | 16,6 | 17 | 18,7 |
| authority#3 | 3 | 10 | 10798 | 16 | 166 | 16,6 | 54 | 10,0 | 1 | 1,1 |
| authority#4 | 2 | 6,7 | 886 | 1,3 | 111 | 11,1 | 36 | 6,7 | 0 | 0 |
| authority#5 | 1 | 3,3 | 6526 | 9,6 | 55 | 5,5 | 18 | 3,3 | 34 | 37,4 |
| authority#6 | 1 | 3,3 | 71 | 0,1 | 55 | 5,5 | 18 | 3,3 | 10 | 11 |
| authority#7 | 0 | 0 | 4106 | 6,1 | 1 | 0,1 | 1 | 0,2 | 0 | 0 |

Table 2: Distribution of examples for the senses of *authority* in different corpora. Proportional and Minimum ratio columns correspond to different ways to apply Semcor bias.

| Sense | 0 | 1 | 2 | 3 | Total web | Semcor |
|---|---|---|---|---|---|---|
| n#church#1 | 0 | 476 | 524 | 0 | 1000 | 60 |
| n#church#2 | 306 | 100 | 561 | 0 | 967 | 58 |
| n#church#3 | 147 | 0 | 20 | 0 | 167 | 10 |
| Overall | 453 | 576 | 1105 | 0 | 453 | 128 |

Table 1: Examples per type (0,1,...) that are acquired for *church* following the Semcor bias, and total examples in Semcor.

| Corpora | Recall |
|---|---|
| Semcor MFS | 47.8 |
| Web - Semcor bias | **49.8** |
| Web - no bias | 38.0 |
| Web - web bias | 39.8 |

Table 3: Performance of the automatically-acquired corpus and effect of bias.

whenever the training and testing samples have different distributions of the senses.

As we are extracting examples automatically, we have to decide how many examples we will be getting for each sense. In order to test the impact of bias, different settings have been tried:

- No bias: we take an equal amount of examples for each sense.
- Web bias: we take all examples gathered from the web.
- Semcor bias: we take a number of examples proportional to the bias of the word in Semcor.

Table 1 shows the number of examples per type (0,1,...) that are acquired for *church* following the Semcor bias. The last column gives the example number in Semcor.

The Semcor bias is not straightforward. In our first approach for Semcor-bias, we assigned 1,000 examples to the major sense in Semcor, and gave the other senses their proportion of examples (when available). But in some cases the distribution of the Semcor bias and that of the actual examples in the web would not fit. The problem is caused when there is not enough examples in the web to fill the expectations of a certain word sense. Table 2 shows shows the different distributions of examples for *authority*, and how the proportional Semcor bias produces a corpus where the percentage of some of the senses is different from that in Semcor, e.g. the first sense only gets 33.7% of the examples, in contrast to the 60% it had in Semcor.

We therefore tried another distribution. We computed, for each word, the minimum ratio of examples that were available for a given target bias and a given number of examples available from the web. We observed that this last approach would reflect better the original bias, and achieve better performance when testing. The minimum-ratio column in Table 2 shows how it approaches much better the proportion of examples in Semcor than that of the more simple proportional approach. The Senseval-testing and Semcor distributions are given; together with the total number of examples in the web, the proportional distribution, and the minimum ratio. There we can see how the distributions of senses in Semcor and Senseval-testing have important differences, although the main sense is the same. It gets clear that if we do not apply minimum ratio the distribution of senses can be affected by the number of examples available.

### 3.3. Local vs. topical features

Previous work on automatic acquisition of examples (Leacock et al., 1998) has reported lower performance when using local collocations formed by PoS tags or closed-class words. In our setting, we observed that local collocations achieved the best precision overall, but the combination of all features obtained the best recall.

However, there were clear differences in the results per word, showing that estimating the best feature-set per word would improve the performance. For the corpus-evaluation experiments, we chose to work with the combination of all features.

## 4. Evaluation

First, we analyzed the impact of bias in the performance of the acquired corpus. The results are shown in Table 3. The precision using the most frequent sense in Semcor (MFS) is also given for reference. The experiment illustrates clearly that when we change the distribution of examples per sense, the performance goes down. The web-corpus with Semcor bias is the only one to beat the baseline. We can see that adding examples in a way that unbalances the sense distribution in training is harmful for the performance.

For our next experiment, we compared the performance using the acquired examples (with Semcor bias and minimum ratio), and the examples from Semcor. We noted that there were clear differences depending on the word-set, and we studied each set separately. The results per word-set are shown in Table 4. The figures correspond to the recall training in Semcor, the web-corpus, and the combination of both.

If we focus in set B (words with less than 10 examples in Semcor), we see that the MFS figure is very low (40.1%). There are even some words that do not have any occurrence in Semcor, and then the sense is chosen at random. It made no sense to train the DL with the handful of examples in Semcor, therefore this result is not in the table. For this set, the bias information from Semcor is also scarce, but the corpus acquired with this information raises the performance to 47.8%.

For set A, the average number of senses is higher, and this raises the results for Semcor MFS (51.9%). We see that the recall for DL training in Semcor is lower that the MFS baseline (50.5%). The main reasons for these low results are the differences between the training and testing corpora (Semcor and Senseval). There has been previous work on portability of hand-tagged corpora that show how some constraints, like the genre or topic of the corpus, affect heavily the results (Martinez and Agirre, 2000). If we train on the web-corpus the results improve, and the the best results are obtained with combination of both corpus , reaching 51.6%. We need to note, however, that it is still lower than the Semcor MFS.

Finally, we will examine the results for the whole set of nouns in the Senseval-2 lexical-sample (last row in Table 4), where we see that the best approach relies on the web-corpus. In order to disambiguate the 29 nouns using only Semcor, we apply MFS when there are less than 10 examples (set B), and train the DLs for the rest.

The results in Table 4 show that the web-corpus raises recall, and the best results are obtained combining the Semcor data and the web examples (50.3%). As we noted, the web-corpus is specially useful when there are few examples in Semcor (set B), therefore we made another test, using the web-corpus only for set B, and applying MFS for set A. The recall was slightly better (50.5%), as is shown in the last column.

| **Word-set** | **MFS** | **Semcor** | **Web** | **Semcor + Web** | **MFS & Web** |
|---|---|---|---|---|---|
| set A (> 10) | **51.9** | 50.5 | 50.9 | 51.6 | **51.9** |
| set B (< 10) | 40.1 | - | 47.7 | **47.8** | **47.8** |
| all words | 47.8 | 47.4 | 49.8 | **50.3** | **50.5** |

Table 4: Recall training in Semcor, the acquired corpus, and a combination of both, compared to that of the Semcor MFS.

## 5. Conclusions and Future Work

This paper explores the large-scale acquisition of sense-tagged examples for WSD. We have used the "monosemous relatives" method to construct automatically a web corpus which, in combination to Semcor, is able to improve the results on the Senseval lexical sample test data. For this, we have shown that the distribution of examples per sense is a critical factor, and we get the best results when using the Semcor distribution. Moreover, we have shown that incorporating information of the Semcor bias, the results we obtained for English are comparable to the use of all-words hand-tagged data (Semcor), which is a very interesting perspective for lesser studied languages. The method can be applied to all the noun senses in WordNet using the data available in `http://ixa2.si.ehu.es/pub/webcorpus` (Agirre and Lopez, 2004).

Still, we only obtained a modest improvement compared to the simple MFS heuristic based on Semcor. We also need to note that the MFS heuristic based on the Senseval-2 training data is some points ahead, indicating that there is still plenty of room for improvement. Future lines of research include refinements of the method to acquire the examples, more powerful Machine Learning methods and exploring feature selection methods for each individual word. Finally, a method to rank automatically word senses according to the target corpus, coupled with the web-corpus, could lead to a fully automatic word-sense disambiguation method.

## 6. Acknowledgments

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

## 7. References

Agirre, E. and O. Lopez, 2004. Publicly available topic signatures for all wordnet nominal senses. *Proceedings of the 4rd International Conference on Languages Resources and Evaluations (LREC).*

Agirre, E. and D. Martinez, 2000. Exploring automatic word sense disambiguation with decision lists and the web. *Procedings of the COLING 2000 Workshop on Semantic Annotation and Intelligent Content*.

Argamon-Engelson, S. and I. Dagan, 1999. Committee-based sample selection for probabilistic classifiers. *Journal of Artificial Intelligence Research*, 11:335–360.

Edmonds, Phil and Scott Cotton, 2001. Senseval-2: Overview. In *Proceedings of the Second International Workshop on evaluating Word Sense Disambiguation Systems*. Toulouse, France.

Leacock, C., M. Chodorow, and G. A. Miller, 1998. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24(1):147–165.

Martinez, D. and E. Agirre, 2000. One sense per collocation and genre/topic variations. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

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

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

Yarowsky, David, 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL).* Cambridge, MA.