Experiment 6.H a): The effect of bias on an automatically built word sense corpus

Document Number	WP6.12
Project ref.	IST-2001-34460
Project Acronym	MEANING
Project full title	Developing Multilingual Web-scale Language Technologies
Project URL	http://www.lsi.upc.es/~nlp/meaning/meaning.html
Availability	Public
Authors: Eneko Agirre (UPV/EHU), David Martinez (UPV/EHU)	

meaning

INFORMATION SOCIETY TECHNOLOGIES

Project ref.	IST-2001-34460
Project Acronym	MEANING
Project full title	Developing Multilingual Web-scale Language Technologies
Security (Distribution level)	Public
Contractual date of delivery	January 2005
Actual date of delivery	January 21, 2005
Document Number	WP6.12
Type	Report
Status & version	v DRAFT
Number of pages	15
WP contributing to the deliberable	WP6
WPTask responsible	German Rigau (UPV/EHU)
Authors	Eneko Agirre (UPV/EHU), David Martinez (UPV/EHU)
Other contributors	
Reviewer	
EC Project Officer	Evangelia Markidou
Authors: Eneko Agirre (UPV/EHU), David Martinez (UPV/EHU)	
Keywords:	
Abstract:	

Contents

1 Introduction

The results of recent WSD exercises, e.g. Senseval-2[1] [Edmonds and Cotton, 2001] show clearly that WSD methods based on hand-tagged examples are the ones performing best. However, the main drawback for supervised WSD is the knowledge acquisition bottleneck: the systems need large amounts of costly hand-tagged data. The situation is more dramatic for lesser studied languages. In order to overcome this problem, different research lines have been explored: automatic acquisition of training examples [Mihalcea, 2002], 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 no specific 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 systems performing best in the English all-words task in Senseval-2 were basically supervised systems trained on Semcor. Unfortunately, for most of the words, this corpus only provides a handful of tagged examples. In fact, only a few systems could overcome the Most Frequent Sense (MFS) baseline, which would tag each word with the sense occurring most frequently in Semcor. In our approach, we will also rely on Semcor as the basic 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. This is a very promising line of research, but one which remains relatively under-studied (cf. Section 2). 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, such as the distribution of the number of training instances per each word sense (bias), and the type of features used for disambiguation (local vs. topical).

Basically, we built three systems, one fully supervised (using examples from both Semcor and automatically acquired examples), one minimally supervised (using the distribution of senses in Semcor and automatically acquired examples) and another fully unsupervised (using an automatically acquired sense rank [McCarthy *et al.*, 2004] and automatically acquired examples).

This paper is structured as follows. First, Section 2 describes previous work on the field. Section 3 introduces the experimental setting for evaluating the acquired corpus. Section4 is devoted to the process of building the corpus, which is evaluated in Section 5. Finally, the conclusions are given in Section 6.

2 Previous work

As we have already mentioned, there is little work on this very promising area. 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 each sense, and

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

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 provided 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. An experiment in the lexical-sample task showed that the method was useful for a subset of the Senseval-2 testing words (results for 5 words are provided).

3 Experimental Setting for Evaluation

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

3.1 Decision Lists

The learning method used to measure the quality of the corpus is **Decision Lists** (DL). This algorithm is described in [Yarowsky, 1994]. In this method, the sense s_k with the highest weighted feature f_i is selected, according to its log-likelihood (see Formula 1). For our implementation, we applied a simple smoothing method: the cases where the denominator is zero are smoothed by the constant 0.1 .

$$weight(s_k, f_i) = \log(\frac{Pr(s_k|f_i)}{\sum_{j \neq k} Pr(s_j|f_i)}) \quad (1)$$

3.2 Features

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

- Local features: Bigrams and trigrams, formed by the word-form, lemma, and part-of-speech[2] 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 using both types together (combination).

[2]The PoS tagging was performed using TnT [Brants, 2000]

3.3 Hand-tagged corpora

Semcor was used as training data for our supervised system. This corpus offers tagged examples for many words, and has been widely used for WSD. It was necessary to use an automatic mapping between the WordNet 1.6 senses in Semcor and the WordNet 1.7 senses in testing [Daude *et al.*, 2000].

For evaluation, the test part of the Senseval-2 English lexical-sample task was chosen. The advantage of this corpus was that we could focus on 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 test examples whose senses were multiwords or phrasal verbs were removed, because they can be efficiently detected with other methods in a preprocess.

It is important to note that the training part of Senseval-2 lexical-sample was not used in the construction of the systems, as our goal was to test the performance we could achieve with minimal resources (i.e. those available for any word). We only relied on the Senseval-2 training bias in preliminary experiments on local/topical features (cf. Table 4), and to serve as a reference for unsupervised performance (cf. Table 5).

3.4 Word-set

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

4 Building the monosemous relatives web corpus

In order to build this corpus[3], 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 (see below). 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 examples based on monosemous relatives for nearly all noun senses in WordNet[4].

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 to 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 sense.

[3]The automatically acquired corpus will be referred indistinctly as web-corpus, or monosemous-corpus
[4]All the examples in this work are publicly available in `http://ixa2.si.ehu.es/pub/sensecorpus`

The monosemous-corpus was built using the simplest technique: we collected examples from the web for each of the monosemous relatives. The relatives have an associated number (type), which correlates roughly with the distance to the target word, and indicates their relevance: the higher the type, the less reliable the relative. A sample of monosemous relatives for different senses of *church*, together with its sense inventory in WordNet 1.7 is shown in Figure 1.

```
Sense 1
church, Christian church, Christianity -- (a group of Christians; any group professing Christian
doctrine or belief)
Sense 2
church, church building -- (a place for public (especially Christian) worship)
Sense 3
church service, church -- (a service conducted in a church)

Monosemous relatives for different senses of church

Synonyms (Type 0):  church building (sense 2), church service (sense 3) ...
Direct hyponyms (Type 1):  Protestant Church (sense 1), Coptic Church (sense 1) ...
Direct hypernyms (Type 2):  house of prayer (sense 2), religious service (sense 3) ...
Distant hyponyms (Type 2,3,4...):  Greek Church (sense 1), Western Church (sense 1)...
Siblings (Type 3):  Hebraism (sense 2), synagogue (sense 2) ...
```

Figure 1: Sense inventory and some monosemous relatives in WordNet 1.7 for *church*.

Distant hyponyms receive a type number equal to the distance to the target sense. Note that we assigned a higher type value to direct hypernyms than to direct hyponyms, as the latter are more useful for disambiguation. We also decided to include siblings, but with a high type value (3).

In the following subsections we will describe step by step the method to construct the corpus. First we will explain the acquisition of the highest possible amount of examples per sense; then we will explain different ways to limit the number of examples per sense for a better performance; finally we will see the effect of training on local or topical features on this kind of corpora.

4.1 Collecting the examples

The examples are collected following these steps

1: We query Google[5] 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 reverse order. This is done because the top hits usually are titles and incomplete sentences that are not so 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: length shorter than 6 words, having more non-alphanumeric characters than words divided by two, or having more words in uppercase than in lowercase.

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

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* for *church*) we can choose not to substitute, because *Protestant*, for instance, 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 on this corpus suggested (although the differences were not significant).

4: For a given word sense, we collect the desired number of examples (see following section) in order of type: we first retrieve all examples of type 0, then type 1, etc. up to type 3 until the necessary examples are obtained. 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 an improvement in future work.

On average, we have acquired roughly 24,000 examples for each of the target words used in this experiment.

4.2 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 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 use 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.
- Automatic ranking: the number of examples is given by a ranking obtained following the method described in [McCarthy *et al.*, 2004]. They used a thesaurus automatically created from the BNC corpus with the method from [Lin, 1998], coupled with WordNet-based similarity measures.
- Semcor bias: we take a number of examples proportional to the bias of the word senses in Semcor.

For example, 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 number of examples in Semcor.

We have to note that the 3 first methods do not require any hand-labeled data, and that the fourth relies in Semcor.

The way to apply the bias is not straightforward in some cases. In our first approach for Semcor-bias, we assigned 1,000 examples to the major sense in Semcor, and gave the other

Sense	0	1	2	3	Total	Semcor
church#1	0	476	524	0	1000	60
church#2	306	100	561	0	967	58
church#3	147	0	20	0	167	10
Overall	453	576	1105	0	2134	128

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

Sense	Semcor		Web corpus								Senseval test	
			Web bias		Semcor Pr		Semcor MR		Automatic MR			
	# ex	%	# ex	%	# ex	%	# ex	%	# ex	%	# ex	%
authority#1	18	60	338	0.5	338	33.7	324	59.9	138	19.3	37	37.4
authority#2	5	16.7	44932	66.4	277	27.6	90	16.6	75	10.5	17	17.2
authority#3	3	10	10798	16	166	16.6	54	10.0	93	13.0	1	1.0
authority#4	2	6.7	886	1.3	111	11.1	36	6.7	67	9.4	0	0
authority#5	1	3.3	6526	9.6	55	5.5	18	3.3	205	28.6	34	34.3
authority#6	1	3.3	71	0.1	55	5.5	18	3.3	71	9.9	10	10.1
authority#7	0	0	4106	6.1	1	0.1	1	0.2	67	9.4	0	0
Overall	30	100	67657	100	1003	100	541	100	716	100	99	100

Table 2: Distribution of examples for the senses of *authority* in different corpora. Pr (proportional) and MR (minimum ratio) columns correspond to different ways to apply Semcor bias.

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 are not enough examples in the web to fill the expectations of a certain word sense.

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 extracted from the web. We observed that this last approach would reflect better the original bias, at the cost of having less examples.

Table 2 presents the different distributions of examples for *authority*. There we can see the Senseval-testing and Semcor distributions, together with the total number of examples in the web; the Semcor proportional distribution (Pr) and minimum ratio (MR); and the automatic distribution. The table illustrates 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 can also see how the distributions of senses in Semcor and Senseval-test have important differences, although the main sense is the same. For the web and automatic distributions, the first sense is different; and in the case of the web distribution, the first hand-tagged sense only accounts for 0.5% of the examples retrieved from the web. Similar distribution discrepancies can be observed for most of the words in the test set. The *Semcor MR* column shows how using minimum ratio we get a better reflection of the proportion of examples in Semcor, compared to the simpler proportional approach (*Semcor Pr*) . For the automatic bias we only used the minimum ratio.

Word	Web bias	Semcor bias	Automatic bias
art	15,387	10,656	2,610
authority	67,657	541	716
bar	50,925	16,627	5,329
bum	17,244	2,555	4,745
chair	24,625	8,512	2,111
channel	31,582	3,235	10,015
child	47,619	3,504	791
church	8,704	5,376	6,355
circuit	21,977	3,588	5,095
day	84,448	9,690	3,660
detention	2,650	1,510	511
dyke	4,210	1,367	843
facility	11,049	8,578	1,196
fatigue	6,237	3,438	5,477
feeling	9,601	1,160	945
grip	20,874	2,209	277
hearth	6,682	1,531	2,730
holiday	16,714	1,248	1,846
lady	12,161	2,959	884
material	100,109	7,855	6,385
mouth	648	287	464
nation	608	594	608
nature	32,553	24,746	9,813
post	34,968	4,264	8,005
restraint	33,055	2,152	2,877
sense	10,315	2,059	2,176
spade	5,361	2,458	2,657
stress	10,356	2,175	3,081
yew	10,767	2,000	8,013
Average	24,137	4,719	3,455
Total	699,086	136,874	100,215

Table 3: Number of examples following different sense distributions. Minimum-ratio is applied for the Semcor and automatic bias.

To conclude this section, Table 3 shows the number of examples acquired automatically following the web bias, the Semcor bias with minimum ratio, and the Automatic bias with minimum ratio.

4.3 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. We performed an early experiment comparing the results using local features, topical features, and a combination of both. In this case we used the web corpus with Senseval training bias, distributed according to the MR approach, and always substituting the target word.

Word	Senseval bias			Semcor bias	Autom. bias
	Loc.	Top.	Comb.		
art	54.2	45.6	47.0	55.6	45.6
authority	47.8	43.2	46.2	41.8	40.0
bar	52.1	55.9	57.2	51.6	26.4
bum	81.2	87.5	85.0	5.0	57.5
chair	88.7	88.7	88.7	88.7	69.4
channel	39.7	53.7	55.9	16.2	30.9
child	56.5	55.6	56.5	54.0	34.7
church	67.7	51.6	54.8	48.4	49.7
circuit	45.3	54.2	56.1	41.5	49.1
day	59.4	54.7	56.8	48.0	12.5
detention	87.5	87.5	87.5	52.1	87.5
dyke	89.3	89.3	89.3	92.9	80.4
facility	28.6	21.4	21.4	26.8	22.0
fatigue	82.5	82.5	82.5	82.5	75.0
feeling	55.1	60.2	60.2	60.2	42.5
grip	19.0	38.0	39.0	16.0	28.2
hearth	73.4	75.0	75.0	75.0	60.4
holiday	96.3	96.3	96.3	96.3	72.2
lady	80.4	73.9	73.9	80.4	23.9
material	43.2	44.2	43.8	54.2	52.3
mouth	36.8	38.6	39.5	54.4	46.5
nation	80.6	80.6	80.6	80.6	80.6
nature	44.4	39.3	40.7	46.7	34.1
post	43.9	40.5	40.5	34.2	47.4
restraint	29.5	37.5	37.1	27.3	31.4
sense	58.1	37.2	38.4	47.7	41.9
spade	74.2	72.6	74.2	67.7	85.5
stress	53.9	46.1	48.7	2.6	27.6
yew	81.5	81.5	81.5	66.7	77.8
Overall	56.5	56.0	57.0	49.8	43.2

Table 4: Recall for all the nouns using the monosemous corpus with Senseval-2 training bias (MR, and substitution), Semcor bias, and Automatic bias. The Senseval-2 results are given by feature type.

The recall (per word and overall) is given in Table 4.

In this setting, we observed that local collocations achieved the best precision overall, but the combination of all features obtained the best recall. The table does not show the precision/coverage figures due to space constraints, but local features achieve 58.5% precision for 96.7% coverage overall, while topical and combination of features have full-coverage.

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.

5 Evaluation

In all experiments, the recall of the systems is presented as evaluation measure. There is total coverage (because of the high overlap of topical features) and the recall and precision are the same[6].

In order to evaluate the acquired corpus, our first task was to analyze the impact of bias. The results are shown in Table 5. There are 2 figures for each distribution: (1) simply assign the first ranked sense, and (2) use the monosemous corpus following the predetermined bias. As we described in Section 3, the testing part of the Senseval-2 lexical sample data was used for evaluation. We also include the results using Senseval2 bias, which is taken from the training part. The recall per word for some distributions can be seen in Table 4.

The results show clearly that when bias information from a hand-tagged corpora is used the recall improves significantly, even when the bias comes from a corpus -Semcor- different from the target corpus -Senseval-. The bias is useful by itself, and we see that the higher the performance of the 1st ranked sense heuristic, the lower the gain using the monosemous corpus. We want to note that in fully unsupervised mode we attain a recall of 43.2% with the automatic ranking. Using the minimally supervised information of bias, we get 49.8% if we have the bias from an external corpus (Semcor) and 57.5% if we have access to the bias of the target corpus (Senseval[7]). This results show clearly that the acquired corpus has useful information about the word senses, and that bias is extremely important.

We will present two further experiments performed with the monosemous corpus resource. The goal of the first will be to measure the WSD performance that we achieve using Semcor as the only supervised data source. In our second experiment, we will compare the performance of our totally unsupervised approach (monosemous corpus and automatic bias) with other unsupervised approaches in the Senseval-2 English lexical task.

5.1 Monosemous corpus and Semcor bias

In this experiment we compared the performance using the monosemous corpus (with Semcor bias and minimum ratio), and the examples from Semcor. We noted that there were clear differences depending on the number of training examples for each word, therefore we studied each word-set described in Section 3.4 separately. The results per word-set are shown in Table 6. The figures correspond to the recall training in Semcor, the web-corpus, and the combination of both.

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

[6]Except for the experiment in Section 4.3, where using local features the coverage is only partial.
[7]Bias obtained from the training-set.

Bias	Type	1st sense	Train examples	Diff.
no bias		18.3	38.0	+19.7
web bias	unsuperv.	33.3	39.8	+6.5
autom. ranking		36.1	43.2	+7.1
Semcor bias	minimally-	47.8	49.8	+2.0
Senseval2 bias	supervised	55.6	57.5	+1.9

Table 5: Performance (recall) on Senseval-2 lexical-sample, using different bias to create the corpus. The *type* column shows the kind of system.

For set A, the average number of examples 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 have been previous works 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 best results are obtained with the combination of both corpora, reaching 51.6%. We need to note, however, that this 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 6), 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 6 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 6: Recall training in Semcor, the acquired web corpus (Semcor bias), and a combination of both, compared to that of the Semcor MFS.

5.2 Monosemous corpus and Automatic bias (unsupervised)

In this experiment we compared the performance of our unsupervised system with other approaches. For this goal, we used the resources available from the Senseval-2 competition[8], where the answers of the participating systems in the different tasks were available. This made possible to compare our results and those of other systems deemed unsupervised by the organizers on the same test data and set of nouns.

From the 5 unsupervised systems presented in the Senseval-2 lexical-sample task as unsupervised, the *WASP-Bench* system relied on lexicographers to hand-code information semi-automatically [Tugwell and Kilgarriff, 2001]. This system does not use the training data, but as it uses manually coded knowledge we think it falls clearly in the supervised category.

The results for the other 4 systems and our own are shown in Table 7. We show the results for the totally unsupervised system and the minimally unsupervised system (Semcor bias). We classified the *UNED* system [Fernandez-Amoros *et al.*, 2001] as minimally supervised. It does not use hand-tagged examples for training, but some of the heuristics that are applied by the system rely on the bias information available in Semcor. The distribution of senses is used to discard low-frequency senses, and also to choose the first sense as a back-off strategy. On the same conditions, our minimally supervised system attains 49.8 recall, nearly 5 points more.

The rest of the systems are fully unsupervised, and they perform significantly worse than our system.

Method	**Type**	**Recall**
Web corpus (Semcor bias)	minimally-	**49.8**
UNED	supervised	45.1
Web corpus (Autom. bias)		**43.3**
Kenneth_Litkowski-clr-ls	unsupervised	35.8
Haynes-IIT2		27.9
Haynes-IIT1		26.4

Table 7: Our minimally supervised and fully unsupervised systems compared to the unsupervised systems (marked in bold) in the 29 noun subset of the Senseval-2 Lexical Sample.

6 Conclusions and Future Work

This paper explores the large-scale acquisition of sense-tagged examples for WSD, which is a very promising line of research, but remains relatively under-studied. We have applied the "monosemous relatives" method to construct automatically a web corpus which we have used to train three systems based on Decision Lists: one fully supervised (applying examples from Semcor and the web corpus), one minimally supervised (relying on the

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

distribution of senses in Semcor and the web corpus) and another fully unsupervised (using an automatically acquired sense rank and the web corpus). Those systems were tested on the Senseval-2 lexical sample test set.

We have shown that the fully supervised system combining our web corpus with the examples in Semcor improves over the same system trained on Semcor alone. This improvement is specially noticeable in the nouns that have less than 10 examples in Semcor. Regarding the minimally supervised and fully unsupervised systems, we have shown that they perform well better than the other systems of the same category presented in the Senseval-2 lexical-sample competition.

The system can be trained for all nouns in WordNet, using the data available at `http://ixa2.si.ehu.es/pub/sensecorpus`.

The research also highlights the importance of bias. Knowing how many examples are to be fed into the machine learning system is a key issue. We have explored several possibilities, and shown that the learning system (DL) is able to learn from the web corpus in all the cases, beating the respective heuristic for sense distribution.

We think that this research opens the opportunity for further improvements. We have to note that the MFS heuristic and the supervised systems based on the Senseval-2 training data are well ahead of our results, and our research aims at investigating ideas to close this gap. Some experiments on the line of adding automatically retrieved examples to available hand-tagged data (Semcor and Senseval-2) have been explored. The preliminary results indicate that this process has to be performed carefully, taking into account the bias of the senses and applying a quality-check of the examples before they are included in the training data.

For the future we also want to test the performance of more powerful Machine Learning methods, explore feature selection methods for each individual word, and more sophisticated ways to combine the examples from the web corpus with those of Semcor or Senseval. Now that the monosemous corpus is available for all nouns, we would also like to test the system on the all-words task. In addition, we will give preference to multiwords that contain the target word when choosing the relatives. Finally, more sophisticated methods to acquire examples are now available, like ExRetriever [Fernandez *et al.*, 2004], and they could open the way to better examples and performance.

7 Acknowledgments

We wish to thank Diana McCarthy, from the University of Sussex, for providing us the sense rank for the target nouns. This research has been partially funded by the European Commission (MEANING IST-2001-34460).

References

[Agirre and Lopez, 2004] E. Agirre and O. Lopez. Publicly available topic signatures for all wordnet nominal senses. In *Proceedings of the 4rd International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, 2004.

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

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

[Brants, 2000] T. Brants. Tnt - a statistical part-of-speech tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference*, Seattle, WA, 2000.

[Daude *et al.*, 2000] J. Daude, L. Padro, and G. Rigau. Mapping wordnets using structural information. In *38th Anual Meeting of the Association for Computational Linguistics (ACL'2000)*, Hong Kong, 2000.

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

[Fernandez-Amoros *et al.*, 2001] D. Fernandez-Amoros, J. Gonzalo, and F. Verdejo. The uned systems at senseval-2. In *Proceedings of the SENSEVAL-2 Workshop. In conjunction with ACL*, Toulouse, France, 2001.

[Fernandez *et al.*, 2004] J. Fernandez, M. Castillo, G. Rigau, J. Atserias, and J. Turmo. Automatic acquisition of sense examples using exretriever. In *Proceedings of the 4rd International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, 2004.

[Leacock *et al.*, 1998] C. Leacock, M. Chodorow, and G. A. Miller. Using corpus statistics and WordNet relations for sense identification. In *Computational Linguistics*, volume 24, pages 147–165, 1998.

[Lin, 1998] D. Lin. Automatic retrieval and clustering of similar words. In *In Proceedings of COLING-ACL*, Montreal, Canada, 1998.

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

[McCarthy *et al.*, 2004] D. McCarthy, R. Koeling, J. Weeds, and J. Carroll. Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL) (to appear)*, Barcelona, Spain, 2004.

[Mihalcea, 2002] R. Mihalcea. Bootstrapping large sense tagged corpora. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Spain, 2002.

[Miller *et al.*, 1993] G. A. Miller, C. Leacock, R. Tengi, and R. Bunker. A semantic concordance. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 303–308, Princeton, NJ, 1993.

[Tugwell and Kilgarriff, 2001] D. Tugwell and A. Kilgarriff. Wasp-bench: a lexicographic tool supporting word sense disambiguation. In *Proceedings of the SENSEVAL-2 Workshop. In conjunction with ACL-2001/EACL-2001*, Toulouse, France, 2001.

[Yarowsky, 1994] D. Yarowsky. Decision lists for lexical ambiguity resolution: Application to accent restoration in spanish and french. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, NM, 1994.

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