Clustering WordNet Word Senses

Eneko Agirre and **Oier Lopez de Lacalle***
Department of computer languages and systems.
University of the Basque Country.
p.k. 649, 20080 Donostia. Spain.
{eneko,jibloleo}@si.ehu.es

Abstract.

This paper presents the results of a set of methods to cluster WordNet word senses. The methods rely on different information sources: confusion matrixes from Senseval-2 Word Sense Disambiguation systems, translation similarities, hand-tagged examples of the target word senses and examples obtained automatically from the web for the target word senses. The clustering results have been evaluated using the coarse-grained word senses provided for the lexical sample in Senseval-2. We have used Cluto, a general clustering environment, in order to test different clustering algorithms. The best results are obtained for the automatically obtained examples, yielding purity values up to 84% on average over 20 nouns.

1 Introduction

WordNet (Miller et al. 1994) is one of the most widely used lexical resources for Natural Language Processing. Among other information, it provides a list of word senses for each word, and has been used in many Word Sense Disambiguation (WSD) systems as the sense inventory of choice. In particular it has been used as the sense inventory for the Senseval-2 English Lexical sample WSD exercise[1] (Edmonds & Cotton 2001).

Many works cite the fine-grainedness of the sense distinctions in WordNet as one of its main problems for practical applications (Dolan 1994; Peters et al. 1998; Resnik & Yarowsky 2000; Mihalcea & Moldovan 2001; Tomuro 2001; Palmer et al. submitted). Senseval-2, for instance, provides both fine-grained (the actual WordNet word senses) and coarse-grained sense distinctions. Figure 1 shows, for instance, the 7 fine-grained senses for the noun channel, and the 4 coarse-grained senses provided by Senseval-2.

There is considerable literature on what makes word senses distinct, but there is no general consensus on which criteria should be followed. Some approaches use an abstraction hierarchy as those found in dictionaries (Kilgarriff 1998), others utilize syntactic patterns such as predicate-argument structure of verbs (Palmer et al., submitted), and others study the word senses from the point of view of systematic polysemy (Peters et al. 1998; Tomuro 2001). From a practical point of view, the need to make two senses distinct will depend on the target application. This is evident for instance in Machine Translation, where some word senses will get the same translation (both *television* and *communication* senses of channel in Figure 1 are translated as *kanal* into Basque) while others don't (*groove* sense of channel is translated as *zirrikitu* in Basque), depending on the target and source languages.

This paper proposes a set of automatic methods to hierarchically cluster the word senses in WordNet. The output of the clustering process is a hierarchical cluster that can be used as complementary information by WSD systems in order to avoid making too fine-grained decisions when not necessary, and return a set of fine-grained senses (a cluster) instead of a single choice. This might allow the error rate of the algorithms to decrease substantially, at the cost of losing discriminative power. Besides the hierarchy complements the flat structure of word senses in WordNet, and could be used by target applications in order to choose the convenient granularity.

The clustering methods that we examine in this paper are based on the following information sources:

1 Similarity matrix for word senses based on the **confusion matrix** of all systems that participated in Senseval-2 (cf. Section 5).
2 Similarity matrix for word senses produced by (Chugur and Gonzalo 2002) using **translation equivalences** in a number of languages (cf. Section 6).
3 **Context of occurrence** for each word sense (cf. Section 7). Two methods were used to derive the contexts: taking them directly from Senseval-2 (hand tagged data), or automatically

* Authors listed in alphabetical order.

[1] In the rest of the paper, the Senseval-2 English Lexical sample exercise will be referred to in short as Senseval-2.

channel	comms	signals	tv	groove	body	passage	water	Synset information: sense number, synonyms, gloss and examples.
water					1	1	1	4. **channel** -- (a deep and relatively narrow body of water (as in a river or a harbor or a strait linking two larger bodies) that allows the best passage for vessels; "the ship went aground in the channel")
passage					1	1	1	2. **channel** -- (a passage for water (or other fluids) to flow through; "the fields were crossed with irrigation channels"; "gutters carried off the rainwater into a series of channels under the street")
body					1	1	1	6. duct, epithelial duct, canal, **channel** -- (a bodily passage or tube lined with epithelial cells and conveying a secretion or other substance; "the tear duct was obstructed"; "the alimentary canal"; "poison is released through a channel in the snake's fangs")
groove				1				3. groove, **channel** -- (a long narrow furrow cut either by a natural process (such as erosion) or by a tool (as e.g. a groove in a phonograph record))
tv		1	1					7. **channel**, television channel, TV channel -- (a television station and its programs; "a satellite TV channel"; "surfing through the channels"; "they offer more than one hundred channels")
signals		1	1					1. **channel**, transmission channel -- (a path over which electrical signals can pass; "a channel is typically what you rent from a telephone company")
comms	1							5. **channel**, communication channel, line -- ((often plural) a means of communication or access; "it must go through official channels"; "lines of communication were set up between the two firms")

Figure 1: The 7 senses for the noun *channel* in WordNet 1.7, and the 4 sense groups given in Senseval-2. The first column shows a mnemonic for each of the word senses, and the last column the sense numbers, words in the synset alongside the glosses and examples in parenthesis. The sense groups are represented by a dendogram (left side of the table) and a similarity matrix (zero values are shown as empty cells).

retrieving them using WordNet information to construct queries over the web (cf. Section 2).

4 Similarity matrix based on the **Topic Signatures** for each word sense (cf. Section 8). The topic signatures were constructed based on the occurrence contexts of the word senses, which, similar to the point above, can be extracted from hand-tagged data or automatically constructed from the Web.

In order to construct the hierarchical clusters we have used Cluto (Karypis 2001), a general clustering environment that has been successfully used in Information Retrieval and Text Categorization. The input to the algorithm can be either a similarity matrix for the word senses of each target word (constructed based on confusion matrixes, translation equivalences, or topic signatures, as mentioned above), or the context of occurrence of each word sense in the form of a vector. Different weighting and clustering schemes were tried (cf. Section 4).

The **gold standard** is based on the manual grouping of word senses provided in Senseval-2. This gold standard is used in order to compute purity and entropy values for the clustering results (cf. Section 9). The number of word sense groups in the gold standard is input to the clustering algorithm as the target number of clusters.

Sections 2 and 3 explain the methods to construct the corpus of word sense examples from the web and the topic signatures respectively. Section 4 presents the clustering environment. Sections 5 through 8 present each of the methods to cluster word senses. Section 9 presents the results of the experiment. Sections 10 and 11 present related work and the conclusions.

2 Retrieving Examples for Word Senses from the Web

Corpora where the occurrences of word senses have been manually tagged are a scarce resource. Semcor (Miller et al. 1994) 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. While being a large corpus, 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 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; Mihalcea and

1. sense: church, Christian_church, Christianity "a group of Christians; any group professing Christian doctrine or belief; "
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 built with the monosemous relatives method to extract examples from the Web. The values in parenthesis correspond to χ^2 values. Only the top scoring terms are shown.

Moldovan 1999; Agirre et al. 2000; Agirre et al. 2001). 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 large corpora or the web. The retrieved examples might not contain the target word, but they contain a word that is closely related to the target word sense.

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[2] 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.

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 marked by three dots in the snippets. 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.

3 Constructing Topic Signatures

Topic signatures try to associate a topical vector to each word sense. The dimensions of these topical vectors are the words in the vocabulary, and the weights try to capture which are the words closer to the target word sense. In other words, each word sense is associated with a set of related words with associated weights.

Figure 2 shows the topic signatures for the three word senses of church. For example, the building sense of church has the terms *house*, *worship*, *building*, *mosque*, etc. with the highest score in the topic signature.

We can build such lists from a sense-tagged corpora just observing which words co-occur distinctively with each sense, or we can try to associate a number of documents from existing

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

corpora to each sense and then analyze the occurrences of words in such documents (cf. previous section).

The method to construct topic signatures proceeds as follows: **(a)** We first organize the documents in collections, one collection per word sense, directly using sense-tagged corpora (e.g. Senseval-2), or exploiting the information in WordNet to build queries and search the web (see section 2). Either way we get one document 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 χ^2. **(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. **(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.

Topic signatures *for words* have been successfully used in summarization tasks (Lin and Hovy 2000). Agirre et al. (2000; 2001) show that it is possible to obtain good quality topic signatures *for word senses*. The topic signatures built in this work can be directly examined in `http://ixa3.si.ehu.es/cgi-bin/signatureak/signaturecgi.cgi`.

4 The Cluto clustering environment

Cluto is a freely available clustering environment that has been successfully used in Information Retrieval and Text Categorization (Karypis 2001; Zhao and Karypis, 2001). The input to the algorithm can be either a similarity matrix for the word senses of each target word (constructed based on confusion matrixes, translation equivalences, or topic signatures, as mentioned above), or the context of occurrence of each word sense in the form of a vector.

In the case of the similarity matrixes the default settings have been tried, as there were only limited possibilities. In the case of using directly the contexts of occurrences, different weighting schemes (plain frequencies, tf.idf), similarity functions (cosine, correlation coefficient) and clustering functions (repeated bisection, optimized repeated bisection, direct, nearest neighbor graph, agglomerative, agglomerative combined with repeated bisection) have been tried (Karypis 2001).

5 Clustering using WSD system confusion matrixes

Given the freely available output of the WSD systems that participated in Senseval-2 (Edmonds & Cotton, 2001), one can construct a confusion matrix for each pair of word senses, constructed as follows:

1. For each pair of word senses (*a,b*), we record the number of times that each WSD system yields *a* when *b* is the correct word sense in the gold standard.
2. This number is divided by the number of occurrences of word sense b and the number of systems.

The rationale is that when the systems confuse two word senses often, we can interpret that the context of the two word senses is similar. For instance, if sense *a* is returned instead of sense *b* always for all systems, the similarity of *a* to *b* will be 1. If the two senses are never confused, then their similarity would be 0. Note that the similarity matrix is not symmetric. Figure 3 shows the similarity matrix for channel based on confusion matrixes, and the resulting hierarchical cluster.

6 Clustering using translation similarities

Chugur and Gonzalo (2002) constructed similarity matrixes for Senseval-2 words using **translation equivalences** in 4 languages, a method proposed by Resnik and Yarowsky (2000). Two word senses are deemed similar if they are often translated with the same word in a given context. More than one language is used to cover as many word sense distinctions as possible. Chugur and Gonzalo kindly provided their similarity matrixes, and we run Cluto directly on them. Figure 4 shows the similarity matrix for channel based on translation similarities, and the resulting hierarchical cluster.

7 Clustering using word sense examples

Clustering of word senses can be cast as a document-clustering problem: each word sense has a pseudo-document associated with it. This pseudo-document is built combining all occurrences of the target word sense (e.g. following the methods in Section 2). Once we have such a pseudo-document for each word sense, we can cluster those documents with the usual techniques, and the output will be clusters of the word senses.

channel	signals	passage	groove	water	comms	body	tv
signals	-	0.00	0.00	0.51	1.03	0.00	3.43
passage	3.95	-	0.17	2.23	0.51	0.51	3.09
groove	0.17	0.00	-	0.00	0.00	0.00	1.20
water	1.72	0.34	0.34	-	0.34	0.00	4.29
comms	0.86	0.51	0.00	2.06	-	0.00	1.72
body	0.51	0.34	0.00	0.17	0.34	-	1.20
tv	2.57	0.00	0.00	0.86	0.69	0.00	-

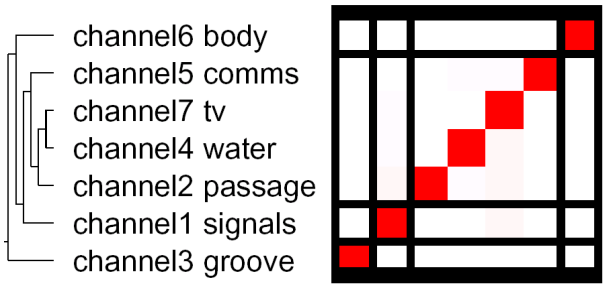

Figure 3: Similarity matrix and clustering results based on the confusion matrix of Senseval 2 systems (units 10^{-3}, note that the matrix is not symmetric). Entropy: 0.429, Purity: 0.714.

channel	signals	passage	groove	water	comms	body	tv
signals	0.88	0.34	0.31	0.21	0.66	0.66	0.71
passage		0.47	0.21	0.35	0.33	0.60	0.30
groove			0.47	0.14	0.34	0.39	0.31
water				0.99	0.16	0.43	0.22
comms					0.88	0.60	0.66
body						1.00	0.67
tv							0.98

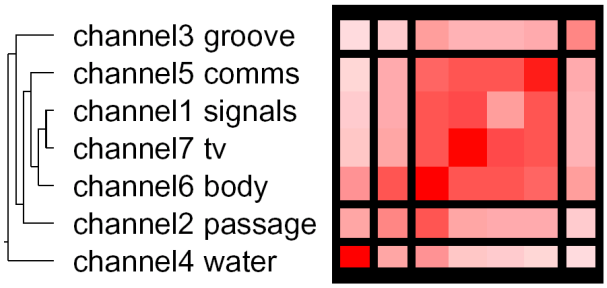

Figure 4: Similarity matrix and clustering results based on multingual method. Entropy: 0.429, Purity: 0.714

channel	signals	passage	groove	water	comms	body	tv
signals	-	0.18	0.04	0.05	0.06	0.29	1.01
passage		-	0.52	0.30	0.26	0.27	1.05
groove			-	0.65	0.20	0.02	0.66
water				-	0.04	0.03	0.00
comms					-	0.45	0.64
body						-	0.55
tv							-

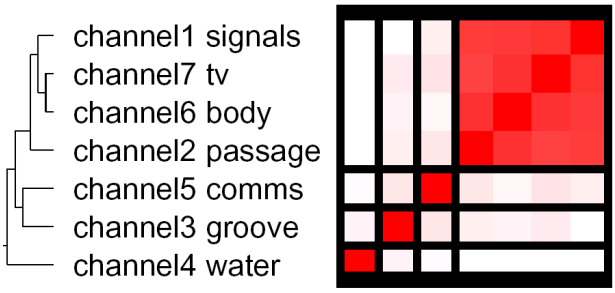

Figure 5: Similarity matrix and resulting hierarchical clusters based on Senseval χ^2 topic signatures for channel (units in 10^{-3}). Entropy: 0.286, Purity: 0.714.

We have used two sources to build the pseudo-documents. On the one hand we have used Senseval 2, using the sentence context only. On the other hand we have used the examples collected from the web, as explained in Section 2. Table 1 shows, among other information, the number of examples available for each of the words considered. A range of clustering parameters was tried on these sets of examples, as explained on Section 4.

8 Clustering using topic signatures

Topic signatures have been constructed for the Senseval 2 nouns using two sets of examples as in the previous section: the Senseval 2 hand-tagged data, and the automatically retrieved sense examples. Once the topic signatures were constructed the similarity matrix was built using the cosine as similarity function among topic signatures. Figure 5 shows the similarity matrix for channel

noun	#senses	#clusters	#senseval	#web	purity
art	4	2	275	23391	0.750
authority	7	4	262	108560	0.571
bar	13	10	360	75792	0.769
bum	4	3	115	25655	1.000
chair	4	2	201	38057	0.750
channel	7	4	181	46493	0.714
child	4	2	189	70416	0.750
circuit	6	4	247	33754	0.833
day	9	5	427	223899	1.000
facility	5	2	172	17878	1.000
fatigue	4	3	116	8596	1.000
feeling	6	4	153	14569	1.000
hearth	3	2	93	10813	0.667
mouth	8	5	171	1585	0.833
nation	4	2	101	1149	1.000
nature	5	3	137	44242	0.600
post	8	5	201	55049	0.625
restraint	6	4	134	49905	0.667
sense	5	4	158	13688	0.800
stress	5	3	112	14528	0.800

Table 1: List of nouns processed in this study. The columns show the number of senses, number of clusters in the gold standard, number of examples in Senseval-2, number of examples retrieved from the web, and best purity value obtained for each word.

based on topic signatures, and the resulting hierarchical cluster.

9 Experiments

In order to evaluate the clustering results we have used as reference gold standard the coarse senses for nouns provided by Senseval-2. This gold standard is used in order to compute purity and entropy values for the clustering results (see below). The number of resulting groups is used as the target number of clusters, that is, in the case of channel (cf. Figure 1) there are 4 sense groups in the gold standard, and therefore Cluto is instructed to build 4 clusters.

Some of the nouns in Senseval-2 had trivial clustering solutions, e.g. when all the word senses form a single cluster, or all clusters are formed by a single word sense. Table 1 shows the 20 nouns that had non-trivial clusters and could therefore be used for evaluation.

The quality of a clustering solution was measured using two different metrics that looked at the gold-standard labels of the word senses assigned to each cluster (Zhao & Karypis 2001). In order to better explain the metrics, we will call the gold-standard sense groups *classes*, as opposed to the clusters returned by the methods. The first metric is the widely used *entropy* measure that looks are how the various classes of word senses are distributed within each cluster, and the second measure is the *purity* that measures the extent to which each cluster contained word senses from primarily one class. A perfect clustering solution will be the one that leads to clusters that contain word senses from only a single class, in which case the entropy will be zero. In general, the smaller the entropy values, the better the clustering solution is.

Method	**entropy**	**purity**
Random	-	0.748
Confusion Matrixes	0.364	0.768
Multilingual Similarity	0.337	0.799
Examples: Senseval (Worse)	0.378	0.744
(Best)	0.338	0.775
Examples: Web (Worse)	0.310	0.800
(Best)	**0.209**	**0.840**
Topic Signature: Senseval	0.286	0.806
Topic Signature: Web	0.358	0.764

Table 2: Entropy and purity results for the different clustering methods.

In a similar fashion, the purity of one cluster is defined to be the fraction of the overall cluster size that the largest class of word senses assigned to that cluster represents. The overall purity of the clustering solution is obtained as a weighted sum of the individual cluster purities. In general, the larger the values of purity, the better the clustering solution is. Evaluation of clustering solutions is not easy, and usually both measures are provided.

As a baseline we built a random baseline, which was computed averaging among 100 random clustering solutions for each word. Each clustering solution was built assigning a cluster to each word sense at random.

Table 2 shows the results in the form of average entropy and purity for each of the clustering methods used. The first line shows the results for the

random baseline. For the confusion matrix and multilingual matrix a single solution is shown. In the case of clustering directly over the examples, the best and worst clustering results are shown for all the combinations of parameters that were tried. Space limits prevent us from showing all combinations. Different results are shown depending on whether the examples were taken from Senseval-2 or from the web examples. Finally, the results over the topic signatures are shown, with different lines depending on the source of the examples.

According to the results, automatically obtained examples proved the best : the optimal combination of clustering parameters gets the best results, and the worst parameter setting gets the third best results. Thus, automatically retrieved examples from the Web are the best source for replicating the gold standard from the alternatives studied in this paper.

If we compare the Web results with the Senseval-2 results (which according to the quality of the hand-tagged data should be better) we see that direct clustering on the examples yields very bad results for Senseval-2 (below random for worst parameter setting). Topic signatures on Senseval-2 data, on the contrary, provide the second best results. We think that the main difference between Senseval-2 and Web data is the amount of data (cf. table 1). It seems that topic signatures provide useful clusters when little data is available, while direct clustering is best for large amounts of data.

We want to note that the random baseline is quite high. Still, the reduction of error for the best result (measured according to purity) is of 35%, i.e. error is reduced from 0.252 to 0.16. We want to note that related literature seldom cite random baselines for clustering problems.

10 Related work

There is some disperse work on word sense clustering. Dolan (1994) proposed a method based on the information on the Machine Readable version of LDOCE (definition text, hierarchies, subcategorization, domains, etc.). The approach is claimed to be useful for improving WSD and mapping different lexical resources at the sense level, but no evaluation is provided. Chen and Chang (1998) use similar methods in order to cluster LDOCE word senses, and they do evaluate their results mapping LDOCE word senses into LLOCE (a thesaurus from the same publishing company) word senses. Results ranging from 100% precision to 60% are reported.

Peters et al. (1998) use the hierarchy of WordNet in order to cluster word senses that are close in the hierarchy. They not only propose a method for clustering, but also try to characterize the relation between similar word senses according to systematic polysemy relations. Mihalcea and Moldovan (2001) present a set of heuristics also based on the structure of WordNet and attain a polysemy reduction of 39% with an error rate of 5.6%. Tomuro (2001) follows a similar approach but introduces the use of more principled algorithms (e.g. the Minimum Description Length) to find systematic polysemy relations between entire WordNet areas. The evaluation is performed comparing to the produced clusters with a gold-standard (WordNet cousins), where 60% precision is attained, and the increase of inter-tagger agreement.

In contrast to approaches based on the structure of WordNet, the output of our proposed methods are based on the distributional similarity among single word senses of a given word, and is able to find individual relations between word senses that cannot be generalized to other word senses. For instance, the gold standard provided (cf. Figure 1) groups together the geographical channel and the bodily channel. We think that both approaches are complementary and plan to further investigate the relation between systematic polysemy, the structure of WordNet and the hierarchical clusters we produce.

In a different research line, Resnik and Yarowsky (2000) propose a multilingual method, which was already outlined in Section 6. The method relies on human translators translating the target word in certain contexts, but could be automated using parallel corpora. Evaluation is conducted comparing the obtained clusters with the sense hierarchies in the Hector dictionary, obtaining 99% correlation. This high correlation contrasts with our poorer results, but unfortunately no random baseline is given for their task. A potential explanation can be the higher quality of the Hector word sense hierarchies, compared to the sense groupings provided by Senseval-2. In fact the similarity matrixes provided by Chugur and Gonzalo (2002), which are based on the same method, attain only 80% purity when compared with our gold standard. We plan to further check the quality of the Senseval-2 groupings, and to explore alternative evaluation methods.

Pantel and Lin (2002) also report related work. They induce soft clusters of words from parsed corpora, using cooccurrence data for words. Each induced cluster represents one concept, and thus words that belong to more than one cluster are polysemous. In order to evaluate their clusters they compare overlap with WordNet clusters with results over 60%.

Finally, Palmer et al. (submitted) provide a deep linguistic analysis of verb senses, giving criteria for manually grouping them.

11 Conclusions and Future work

This paper has presented the results of a set of methods to cluster WordNet word senses. The methods rely on different information sources: confusion matrixes from Senseval-2 systems, translation similarities, hand-tagged examples of the target word senses and examples obtained automatically from the web for the target word senses. The clustering results have been evaluated using the coarse-grained word senses provided for the lexical sample in Senseval-2. We have used Cluto, a general clustering environment, in order to test different clustering algorithms.

The best results are obtained with the automatically obtained examples, with purity values up to 84% on average over 20 nouns.

We are currently acquiring 1,000 snippets from Google for each monosemous noun in WordNet. The total amount of monosemous nouns in WN1.6 is 90,645 and we have currently acquired examples for 98% of them. The examples take nearly 16 Gigabytes. We plan to provide word sense clusters of comparable quality for all nouns in WordNet soon.

We also plan to perform a task based evaluation, using the hierarchical clusters to guide a WSD algorithm, and to further investigate the relation of the produced clusters and studies of systematic polysemy.

Acknowledgements

We want to thank Irina Chugur and Julio Gonzalo for kindly providing us with the multilingual data, and the three anonymous reviewers for their valuable comments. This work has been funded by the European Commision (project MEANING IST-2001-34460) and the MCYT (project HERMES 2000-0335-C03-03).

References

(Agirre et al. 2000) Agirre, E., Ansa, O., Martinez, D. and E. Hovy. Enriching very large ontologies with topic signatures. *Proc. of the workshop on Ontology Learning held in conjunction with ECAI*. 2000.

(Agirre et al. 2001) Agirre, E., Ansa, O., Martinez, D. and E. Hovy. Enriching WordNet concepts with topic signatures. *Proceedings of the SIGLEX workshop on "WordNet and Other Lexical Resources: Applications, Extensions and Customizations" held in conjunction with NAACL*. Pittsburgh, USA. 2001.

(Chen & Chang 1998) Chen, J.N. and J.S. Chang. Topical Clustering of MRD Senses Based on Information Retrieval Techniques. *Computational Linguistics* 24(1): 61-95. 1998.

(Chugur & Gonzalo 2002) Chugur, I and J. Gonzalo. A study of Polysemy and Sense Proximity in the Senseval-2 test suite. *Proc of the ACL Workshop: Word Sense Disambiguation: recent successes and future directions*. 2002.

(Dolan 1994) Dolan, W. 1994 Word Sense Ambiguation: Clustering Related Senses, *Proc. 15th Int'l. Conf. Computational Linguistics, ACL*, Morristown, N.J., 1994, pp. 712-716.

(Edmonds & Cotton 2001) Edmonds, P. and S. Cotton (eds.). *Proceedings of the Senseval-2 Workshop.* Association for Computational Linguistics. Toulousse, France. 2001. See also http://www.sle.sharp.co.uk/senseval2/

(Karypis 2001) Karypis, G. *CLUTO .A Clustering Toolkit.* Technical Report**:** #02-01 Department of Computer Science Minneapolis, MN 55455, University of Minnesota. 2001. Also available in http://www.cs.umn.edu/~karypis.

(Kilgarriff 1998) Kilgarriff, A. Inter-tagger agreement. In *Advanced papers of the SENSEVAL Workshop.* Sussex, UK.

(Leacock et al. 1998) Leacock, C., Chodorow, M. and Miller, G.A. *Using Corpus Statistics and WordNet Relations for Sense Identication*. Computational Linguistics, 24(1):147-166, March 1998.

(Lin & Hovy 2000) Lin, C.-Y. and E.H. Hovy. The Automated Acquisition of Topic Signatures for Text Summarization. *Proc. of the COLING Conference.* Strasbourg, France. August, 2000.

(Mihalcea & Moldovan 1999) Mihalcea, R. and D.I. Moldovan. An Automatic Method for Generating Sense Tagged Corpora. *Proc. of the Conference of the American Association of Artificial Intelligence*. 1999.

(Mihalcea & Moldovan 2001) Mihalcea, R. and D.I. Moldovan. Automatic generation of a coarse grained WordNet. *Proceedings of the SIGLEX workshop on "WordNet and Other Lexical Resources: Applications, Extensions and Customizations" held in conjunction with NAACL*. Pittsburgh, USA. 2001.

(Miller et al. 1990) Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Five Papers on WordNet. *Special Issue of International Journal of Lexicography*, 3(4). 1990.

(Palmer et al. submitted) Palmer, M., Trang, H.T. and C. Fellbaum. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. Submitted to *Natural Language Engineering*.

(Pantel & Lin 2002) Pantel P. and D. Lin. Discovering Word Senses from Text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining* 2002. pp. 613-619. Edmonton, Canada, 2002.

(Peters et al. 1998) Peters, W., Peters, I, and P. Vossen. Automatic Sense Clustering in EuroWordNet. *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada,, Spain, pp. 409-416, 1998.

(Resnik & Yarowsky 2000) Resnik, P. and D. Yarowsky. Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation, *Natural Language Engineering* 5(2), pp. 113-133, 2000.

(Tomuro 2001) Tomuro, N. Tree-cut and A Lexicon based on Systematic Polysemy. *Proceedings of the meeting of the North American Association for Computational Linguistics*. Pittsburgh, USA. 2001.

(Zhao & Karypis 2001) Ying Zhao and George Karypis. *Criterion functions for document clustering: Experiments and analysis.* Technical Report TR #01–40, Department of Computer Science, University of Minnesota, Minneapolis, USA, 2001.