SemEval-2007 Task 07: Coarse-Grained English All-Words Task

Roberto Navigli
Università di Roma "La Sapienza"
Dipartimento di Informatica
Via Salaria, 00198 - Roma Italy
navigli@di.uniroma1.it

Kenneth C. Litkowski
CL Research
9208 Gue Road
Damascus MD 20872
ken@clres.com

Orin Hargraves
Lexicographer
orinhargraves
@googlemail.com

Abstract

This paper presents the coarse-grained English all-words task at SemEval-2007. We describe our experience in producing a coarse version of the WordNet sense inventory and preparing the sense-tagged corpus for the task. We present the results of participating systems and discuss future directions.

1 Introduction

It is commonly thought that one of the major obstacles to high-performance Word Sense Disambiguation (WSD) is the fine granularity of sense inventories. State-of-the-art systems attained a disambiguation accuracy around 65% in the Senseval-3 all-words task (Snyder and Palmer, 2004), where WordNet (Fellbaum, 1998) was adopted as a reference sense inventory. Unfortunately, WordNet is a fine-grained resource, encoding sense distinctions that are difficult to recognize even for human annotators (Edmonds and Kilgarriff, 2002). Making WSD an enabling technique for end-to-end applications clearly depends on the ability to deal with reasonable sense distinctions.

The aim of this task was to explicitly tackle the granularity issue and study the performance of WSD systems on an all-words basis when a coarser set of senses is provided for the target words. Given the need of the NLP community to work on freely available resources, the solution of adopting a different computational lexicon is not viable. On the other hand, the production of a coarse-grained sense inventory is not a simple task. The main issue is certainly the subjectivity of sense clusters. To overcome this problem, different strategies can be adopted. For instance, in the OntoNotes project (Hovy et al., 2006) senses are grouped until a 90% inter-annotator agreement is achieved. In contrast, as we describe in this paper, our approach is based on a mapping to a previously existing inventory which encodes sense distinctions at different levels of granularity, thus allowing to induce a sense clustering for the mapped senses.

We would like to mention that another SemEval-2007 task dealt with the issue of sense granularity for WSD, namely Task 17 (subtask #1): Coarse-grained English Lexical Sample WSD. In this paper, we report our experience in organizing Task 07.

2 Task Setup

The task required participating systems to annotate open-class words (i.e. nouns, verbs, adjectives, and adverbs) in a test corpus with the most appropriate sense from a coarse-grained version of the WordNet sense inventory.

2.1 Test Corpus

The test data set consisted of 5,377 words of running text from five different articles: the first three (in common with Task 17) were obtained from the WSJ corpus, the fourth was the Wikipedia entry for *computer programming*[1], the fifth was an excerpt of Amy Steedman's *Knights of the Art*, biographies of Italian painters[2]. We decided to add the last two

[1] http://en.wikipedia.org/wiki/Computer_programming
[2] http://www.gutenberg.org/etext/529

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 30–35,
Prague, June 2007.

article	domain	words	annotated
d001	JOURNALISM	951	368
d002	BOOK REVIEW	987	379
d003	TRAVEL	1311	500
d004	COMPUTER SCIENCE	1326	677
d005	BIOGRAPHY	802	345
total		5377	2269

Table 1: Statistics about the five articles in the test data set.

texts to the initial dataset as we wanted the corpus to have a size comparable to that of previous editions of all-words tasks.

In Table 1 we report the domain, number of running words, and number of annotated words for the five articles. We observe that articles d003 and d004 are the largest in the corpus (they constitute 51.87% of it).

2.2 Creation of a Coarse-Grained Sense Inventory

To tackle the granularity issue, we produced a coarser-grained version of the WordNet sense inventory[3] based on the procedure described by Navigli (2006). The method consists of automatically mapping WordNet senses to top level, numbered entries in the Oxford Dictionary of English (ODE, (Soanes and Stevenson, 2003)). The semantic mapping between WordNet and ODE entries was obtained in two steps: first, we disambiguated with the SSI algorithm (Navigli and Velardi, 2005) the definitions of the two dictionaries, together with additional information (hypernyms and domain labels); second, for each WordNet sense, we determined the best matching ODE coarse entry. As a result, WordNet senses mapped to the same ODE entry were assigned to the same sense cluster. WordNet senses with no match were associated with a singleton sense.

In contrast to the automatic method above, the sense mappings for all the words in our test corpus were manually produced by the third author, an expert lexicographer, with the aid of a mapping interface. Not all the words in the corpus could be mapped directly for several reasons: lacking entries in ODE (e.g. adjectives underlying and shivering), different spellings (e.g. after-effect vs. aftereffect, halfhearted vs. half-hearted, etc.), derivatives (e.g. procedural, gambler, etc.). In most of the cases, we asked the lexicographer to map senses of the original word to senses of lexically-related words (e.g. WordNet senses of procedural were mapped to ODE senses of procedure, etc.). When this mapping was not straightforward, we just adopted the WordNet sense inventory for that word.

We released the entire sense groupings (those induced from the manual mapping for words in the test set plus those automatically derived on the other words) and made them available to the participants.

2.3 Sense Annotation

All open-class words (i.e. nouns, verbs, adjectives, and adverbs) with an existing sense in the WordNet inventory were manually annotated by the third author. Multi-word expressions were explicitly identified in the test set and annotated as such (this was made to allow a fair comparison among systems independent of their ability to identify multi-word expressions).

We excluded auxiliary verbs, uncovered phrasal and idiomatic verbs, exclamatory uses, etc. The annotator was allowed to tag words with multiple coarse senses, but was asked to make a single sense assignment whenever possible.

The lexicographer annotated an overall number of 2,316 content words. 47 (2%) of them were excluded because no WordNet sense was deemed appropriate. The remaining 2,269 content words thus constituted the test data set. Only 8 of them were assigned more than one sense: specifically, two coarse senses were assigned to a single word instance[4] and two distinct fine-grained senses were assigned to 7 word instances. This was a clear hint that the sense clusters were not ambiguous for the vast majority of words.

In Table 2 we report information about the polysemy of the word instances in the test set. Overall, 29.88% (678/2269) of the word instances were monosemous (according to our coarse sense inventory). The average polysemy of the test set with the coarse-grained sense inventory was 3.06 compared to an average polysemy with the WordNet inventory

[3] We adopted WordNet 2.1, available from: http://wordnet.princeton.edu

[4] d005.s004.t015

polysemy	N	V	A	R	all
monosemous	358	86	141	93	678
polysemous	750	505	221	115	1591
total	1108	591	362	208	2269

Table 2: Statistics about the test set polysemy (N = nouns, V = verbs, A = adjectives, R = adverbs).

of 6.18.

2.4 Inter-Annotator Agreement

Recent estimations of the inter-annotator agreement when using the WordNet inventory report figures of 72.5% agreement in the preparation of the English all-words test set at Senseval-3 (Snyder and Palmer, 2004) and 67.3% on the Open Mind Word Expert annotation exercise (Chklovski and Mihalcea, 2002).

As the inter-annotator agreement is often considered an upper bound for WSD systems, it was desirable to have a much higher number for our task, given its coarse-grained nature. To this end, beside the expert lexicographer, a second author independently performed part of the manual sense mapping (590 word senses) described in Section 2.2. The pairwise agreement was 86.44%.

We repeated the same agreement evaluation on the sense annotation task of the test corpus. A second author independently annotated part of the test set (710 word instances). The pairwise agreement between the two authors was 93.80%. This figure, compared to those in the literature for fine-grained human annotations, gives us a clear indication that the agreement of human annotators strictly depends on the granularity of the adopted sense inventory.

3 Baselines

We calculated two baselines for the test corpus: a *random baseline*, in which senses are chosen at random, and the *most frequent baseline* (MFS), in which we assign the first WordNet sense to each word in the dataset.

Formally, the accuracy of the random baseline was calculated as follows:

$$BL_{Rand} = \frac{1}{|T|}\sum_{i=1}^{|T|}\frac{1}{|CoarseSenses(w_i)|}$$

where T is our test corpus, w_i is the i-th word instance in T, and $CoarseSenses(w_i)$ is the set of coarse senses for w_i according to the sense clustering we produced as described in Section 2.2.

The accuracy of the MFS baseline was calculated as:

$$BL_{MFS} = \frac{1}{|T|}\sum_{i=1}^{|T|}\delta(w_i, 1)$$

where $\delta(w_i, k)$ equals 1 when the k-th sense of word w_i belongs to the cluster(s) manually associated by the lexicographer to word w_i (0 otherwise). Notice that our calculation of the MFS is based on the frequencies in the SemCor corpus (Miller et al., 1993), as we exploit WordNet sense rankings.

4 Results

12 teams submitted 14 systems overall (plus two systems from a 13^{th} withdrawn team that we will not report). According to the SemEval policy for task organizers, we remark that the system labelled as UoR-SSI was submitted by the first author (the system is based on the Structural Semantic Interconnections algorithm (Navigli and Velardi, 2005) with a lexical knowledge base composed by WordNet and approximately 70,000 relatedness edges). Even though we did not specifically enrich the algorithm's knowledge base on the task at hand, we list the system separately from the overall ranking.

The results are shown in Table 3. We calculated a MFS baseline of 78.89% and a random baseline of 52.43%. In Table 4 we report the F1 measures for all systems where we used the MFS as a backoff strategy when no sense assignment was attempted (this possibly reranked 6 systems - marked in bold in the table - which did not assign a sense to all word instances in the test set). Compared to previous results on fine-grained evaluation exercises (Edmonds and Kilgarriff, 2002; Snyder and Palmer, 2004), the systems' results are much higher. On the other hand, the difference in performance between the MFS baseline and state-of-the-art systems (around 5%) on coarse-grained disambiguation is comparable to that of the Senseval-3 all-words exercise. However, given the novelty of the task we believe that systems can achieve even better perfor-

System	A	P	R	F1
NUS-PT	100.0	82.50	82.50	82.50
NUS-ML	100.0	81.58	81.58	81.58
LCC-WSD	100.0	81.45	81.45	81.45
GPLSI	100.0	79.55	79.55	79.55
BL_{MFS}	100.0	78.89	78.89	78.89
UPV-WSD	100.0	78.63	78.63	78.63
TKB-UO	100.0	70.21	70.21	70.21
PU-BCD	90.1	69.72	62.80	66.08
RACAI-SYNWSD	100.0	65.71	65.71	65.71
SUSSX-FR	72.8	71.73	52.23	60.44
USYD	95.3	58.79	56.02	57.37
UOFL	92.7	52.59	48.74	50.60
SUSSX-C-WD	72.8	54.54	39.71	45.96
SUSSX-CR	72.8	54.30	39.53	45.75
UOR-SSI†	100.0	83.21	83.21	83.21

Table 3: System scores sorted by F1 measure (A = attempted, P = precision, R = recall, F1 = F1 measure, †: system from one of the task organizers).

System	F1
NUS-PT	82.50
NUS-ML	81.58
LCC-WSD	81.45
GPLSI	79.55
BL_{MFS}	78.89
UPV-WSD	78.63
SUSSX-FR	77.04
TKB-UO	70.21
PU-BCD	69.72
RACAI-SYNWSD	65.71
SUSSX-C-WD	64.52
SUSSX-CR	64.35
USYD	58.79
UOFL	54.61
UOR-SSI†	83.21

Table 4: System scores sorted by F1 measure with MFS adopted as a backoff strategy when no sense assignment is attempted (†: system from one of the task organizers). Systems affected are marked in bold.

mance by heavily exploiting the coarse nature of the sense inventory.

In Table 5 we report the results for each of the five articles. The interesting aspect of the table is that documents from some domains seem to have predominant senses different from those in SemCor. Specifically, the MFS baseline performs more poorly on documents d004 and d005, from the COMPUTER SCIENCE and BIOGRAPHY domains respectively. We believe this is due to the fact that these documents have specific predominant senses, which correspond less often to the most frequent sense in SemCor than for the other three documents. It is also interesting to observe that different systems perform differently on the five documents (we highlight in bold the best performing systems on each article).

Finally, we calculated the systems' performance by part of speech. The results are shown in Table 6. Again, we note that different systems show different performance depending on the part-of-speech tag. Another interesting aspect is that the performance of the MFS baseline is very close to state-of-the-art systems for adjectives and adverbs, whereas it is more than 3 points below for verbs, and around 5 for nouns.

System	N	V	A	R
NUS-PT	**82.31**	**78.51**	**85.64**	89.42
NUS-ML	81.41	78.17	82.60	**90.38**
LCC-WSD	80.69	78.17	85.36	87.98
GPLSI	80.05	74.45	82.32	86.54
BL_{MFS}	77.44	75.30	84.25	87.50
UPV-WSD	79.33	72.76	84.53	81.25
TKB-UO	70.76	62.61	78.73	74.04
PU-BCD	71.41	59.69	66.57	55.67
RACAI-SYNWSD	64.02	62.10	71.55	75.00
SUSSX-FR	68.09	51.02	57.38	49.38
USYD	56.06	60.43	58.00	54.31
UOFL	57.65	48.82	25.87	60.80
SUSSX-C-WD	52.18	35.64	42.95	46.30
SUSSX-CR	51.87	35.44	42.95	46.30
UOR-SSI†	84.12	78.34	85.36	88.46

Table 6: System scores by part-of-speech tag (N = nouns, V = verbs, A = adjectives, R = adverbs) sorted by overall F1 measure (best scores are marked in bold, †: system from one of the task organizers).

	d001		d002		d003		d004		d005	
System	P	R	P	R	P	R	P	R	P	R
NUS-PT	**88.32**	**88.32**	88.13	88.13	**83.40**	**83.40**	76.07	76.07	**81.45**	**81.45**
NUS-ML	86.14	86.14	**88.39**	**88.39**	81.40	81.40	**76.66**	**76.66**	79.13	79.13
LCC-WSD	87.50	87.50	87.60	87.60	81.40	81.40	75.48	75.48	80.00	80.00
GPLSI	83.42	83.42	86.54	86.54	80.40	80.40	73.71	73.71	77.97	77.97
BL_{MFS}	85.60	85.60	84.70	84.70	77.80	77.80	75.19	75.19	74.20	74.20
UPV-WSD	84.24	84.24	80.74	80.74	76.00	76.00	77.11	77.11	77.10	77.10
TKB-UO	78.80	78.80	72.56	72.56	69.40	69.40	70.75	70.75	58.55	58.55
PU-BCD	77.16	67.94	75.52	67.55	64.96	58.20	68.86	61.74	64.42	60.87
RACAI-SYNWSD	71.47	71.47	72.82	72.82	66.80	66.80	60.86	60.86	59.71	59.71
SUSSX-FR	79.10	57.61	73.72	53.30	74.86	52.40	67.97	48.89	65.20	51.59
USYD	62.53	61.69	59.78	57.26	60.97	57.80	60.57	56.28	47.15	45.51
UOFL	61.41	59.24	55.93	52.24	48.00	45.60	53.42	47.27	44.38	41.16
SUSSX-C-WD	66.42	48.37	61.31	44.33	55.14	38.60	50.72	36.48	42.13	33.33
SUSSX-CR	66.05	48.10	60.58	43.80	59.14	41.40	48.67	35.01	40.29	31.88
UOR-SSI†	86.14	86.14	85.49	85.49	79.60	79.60	86.85	86.85	75.65	75.65

Table 5: System scores by article (best scores are marked in bold, †: system from one of the task organizers).

5 Systems Description

In order to allow for a critical and comparative inspection of the system results, we asked the participants to answer some questions about their systems. These included information about whether:

1. the system used semantically-annotated and unannotated resources;
2. the system used the MFS as a backoff strategy;
3. the system used the coarse senses provided by the organizers;
4. the system was trained on some corpus.

We believe that this gives interesting information to provide a deeper understanding of the results. We summarize the participants' answers to the questionnaires in Table 7. We report about the use of semantic resources as well as semantically annotated corpora (SC = SemCor, DSO = Defence Science Organisation Corpus, SE = Senseval corpora, OMWE = Open Mind Word Expert, XWN = eXtended WordNet, WN = WordNet glosses and/or relations, WND = WordNet Domains), as well as information about the use of unannotated corpora (UC), training (TR), MFS (based on the SemCor sense frequencies), and the coarse senses provided by the organizers (CS). As expected, several systems used lexico-semantic information from the WordNet semantic network and/or were trained on the SemCor semantically-annotated corpus.

Finally, we point out that all the systems performing better than the MFS baseline adopted it as a backoff strategy when they were not able to output a sense assignment.

6 Conclusions and Future Directions

It is commonly agreed that Word Sense Disambiguation needs emerge and show its usefulness in end-to-end applications: after decades of research in the field it is still unclear whether WSD can provide a relevant contribution to real-world applications, such as Information Retrieval, Question Answering, etc. In previous Senseval evaluation exercises, state-of-the-art systems achieved performance far below 70% and even the agreement between human annotators was discouraging. As a result of the discussion at the Senseval-3 workshop in 2004, one of the aims of SemEval-2007 was to tackle the problems at the roots of WSD. In this task, we dealt with the granularity issue which is a major obstacle to both system and human annotators. In the hope of overcoming the current performance upper bounds, we

System	SC	DSO	SE	OMWE	XWN	WN	WND	OTHER	UC	TR	MFS	CS
GPLSI	√	×	√	×	×	√	×	×	×	√	√	√
LCC-WSD	√	×	√	√	√	√	×	×	×	√	√	√
NUS-ML	√	×	×	×	×	×	×	×	√	√	√	×
NUS-PT	√	√	×	×	×	×	×	Parallel corpus	×	√	√	√
PU-BCD	√	×	×	×	×	×	×	×	×	√	×	√
RACAI-SYNWSD	×	×	×	×	×	√	√	×	√	×	×	√
SUSSX-C-WD	×	×	×	×	×	×	×	×	√	×	×	×
SUSSX-CR	×	×	×	×	×	×	×	×	√	×	×	×
SUSSX-FR	×	×	×	×	×	×	×	×	√	×	×	√
TKB-UO	×	×	×	×	×	√	×	×	×	×	×	×
UOFL	×	×	×	×	√	√	×	×	×	×	×	×
UOR-SSI†	×	×	×	×	×	√	×	SSI LKB	×	×	√	×
UPV-WSD	×	×	×	×	×	√	√	×	×	×	√	×
USYD	√	×	√	×	×	√	×	×	√	√	√	√

Table 7: Information about participating systems (SC = SemCor, DSO = Defence Science Organisation Corpus, SE = Senseval corpora, OMWE = Open Mind Word Expert, XWN = eXtended WordNet, WN = WordNet glosses and/or relations, WND = WordNet Domains, UC = use of unannotated corpora, TR = use of training, MFS = most frequent sense backoff strategy, CS = use of coarse senses from the organizers, †: system from one of the task organizers).

proposed the adoption of a coarse-grained sense inventory. We found the results of participating systems interesting and stimulating. However, some questions arise. First, it is unclear whether, given the novelty of the task, systems really achieved the state of the art or can still improve their performance based on a heavier exploitation of coarse- and fine-grained information from the adopted sense inventory. We observe that, on a technical domain such as computer science, most supervised systems performed worse due to the nature of their training set. Second, we still need to show that coarse senses can be useful in real applications. Third, a full coarse sense inventory is not yet available: this is a major obstacle to large-scale *in vivo* evaluations. We believe that these aspects deserve further investigation in the years to come.

Acknowledgments

This work was partially funded by the Interop NoE (508011), 6^{th} European Union FP. We would like to thank Martha Palmer for providing us the first three texts of the test corpus.

References

Tim Chklovski and Rada Mihalcea. 2002. Building a sense tagged corpus with open mind word expert. In *Proc. of ACL 2002 Workshop on WSD: Recent Successes and Future Directions*. Philadelphia, PA.

Philip Edmonds and Adam Kilgarriff. 2002. Introduction to the special issue on evaluating word sense disambiguation systems. *Journal of Natural Language Engineering*, 8(4):279–291.

Christiane Fellbaum, editor. 1998. *WordNet: an Electronic Lexical Database*. MIT Press.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Comp. Volume*, pages 57–60, New York City, USA.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 303–308, Princeton, NJ, USA.

Roberto Navigli and Paola Velardi. 2005. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(7):1063–1074.

Roberto Navigli. 2006. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proc. of the 44th Annual Meeting of the Association for Computational Linguistics joint with the 21st International Conference on Computational Linguistics (COLING-ACL 2006)*, pages 105–112. Sydney, Australia.

Benjamin Snyder and Martha Palmer. 2004. The english all-words task. In *Proc. of ACL 2004 SENSEVAL-3 Workshop*, pages 41–43. Barcelona, Spain.

Catherine Soanes and Angus Stevenson, editors. 2003. *Oxford Dictionary of English*. Oxford University Press.