SENSEVAL-3: Third International Workshop on the Evaluation of Systems
for the Semantic Analysis of Text, Barcelona, Spain, July 2004
Association for Computational Linguistics

The English All-Words Task

Benjamin Snyder and Martha Palmer

University of Pennsylvania

{bsnyder3,mpalmer@}linc.cis.upenn.edu

Abstract

We describe our experience in preparing the sense-tagged corpus used in the English all-words task and we tabulate the scores.

1 Test Corpus

The test data consisted of approximately 5,000 words of running text from two Wall Street Journal articles and one excerpt from the Brown Corpus. The three texts represent three distinct domains: editorial, news story and fiction.[1] They were culled from the Penn Treebank II.

All verbs, nouns, and adjectives were double-annotated with WordNet 1.7.1 (Fellbaum, 1998) senses, and then adjudicated and corrected by a third person.[2] The annotators were encouraged to indicate multi-word constructions when WordNet contains an appropriate entry. The annotators were allowed to tag words with multiple senses, but were asked to pick a single sense whenever possible. The annotators were also asked to indicate when no sense in WordNet fits the meaning of the word (marked as U).

A total of 2,212 words were tagged. Because some of these words were part of multi-word constructions, the total number of answers was 2,081.[3] There were an average of 1.03 senses per answer after adjudication.

A SENSEVAL-style corpus, indicating head words to be tagged along with satellite words for multi-word expressions, was created and distributed to the participants along with the original syntactic and part-of-speech annotation from the Treebank II files. The participants were given one week to run their systems on the test-data and submit the results.

2 Inter-annotator Agreement

The inter-annotator agreement rate in the preparation of the corpus was approximately 72.5%. Verbs had the lowest agreement rate at 67.8%, followed by nouns at 74.9% and adjectives at 78.5%.

The disagreements tended to cluster around a relatively small group of difficult words. Only 38% of all word types and 57% of word types with more than five tokens had any disagreement at all.

One word with very low agreement was the adjective *national*. In six out of seven instances one annotator chose sense two:

limited to or in the interests of a particular nation

while the other annotator chose sense three:

concerned with or applicable to or belonging to an entire nation or country

The remaining five senses were never used. The main difference between these two senses seems to be that the former applies when the use of *national* is intended to draw a contrast with something *international*, and the latter applies when *national* is intended to draw a contrast with something *local*. Two points about this should be made: (a) these two senses are closely related and in actual uses of the word it may be impossible to judge which of them is most applicable; (b) the actual distinction between the two senses had to be inferred from

[1] the news story, wsj_1778, mostly consists of excerpts from electronic bulletin boards in the wake of the 1989 San Francisco earthquake. The editorial is wsj_1695, and the fiction excerpt is cl23.

[2] The annotators and adjudicators all had previous experience doing WordNet sense tagging and all have advanced degrees in either computational linguistics or theoretical linguistics.

[3] Due to various reasons, only 2,041 of these were used in the scoring of the systems. The 40 removed instances include auxiliaries, words which do not have WordNet entries, instances with incorrect TreeBank part-of-speech tags, and instances where the test-data had been formatted incorrectly

the glosses. The glosses do not themselves make the sense distinctions explicit.

In fact, we believe that most of the annotator disagreements were, like this example, between closely related WordNet senses with only subtle (and often inexplicit) distinctions and that more coarse-grained sense distinctions are needed (Palmer et al., 2004).

3 Systems and Scores

26 systems were submitted by a total of 16 teams. The system names, along with email contacts are listed in table 3. Two sets of scores were computed for each system.

For the first set of scores ("With U"), we assumed an answer of U (untaggable) whenever the system failed to provide a sense. Thus the instance would be scored as correct if the answer key also marked it as U, and incorrect otherwise.

For the second set of scores ("Without U"), we simply skipped every instance where the system did not provide a sense. Thus precision was not affected by those instances, but recall was lowered.

Even though any given team may have intended their results to be interpreted one way or the other, we have included both sets of scores for comparative purposes. Table 1 shows the system performance under the first interpretation of the results ("With U"). The average precision and recall is 52.2%.

Table 2 shows the system performance under the second interpretation of the results ("Without U"). The average precision is 57.4% and 51.9% is the average recall.

Since comprehensive groupings of the WordNet senses do not yet exist, all results given are the result of fine-grained scoring.

Although we did not compute a baseline score, we received several baseline figures from our participants. Deniz Yuret, of Koc University, computed a baseline of 60.9% precision and recall by using the first WordNet entry for the given word and part-of-speech. Bart Decadt, of the University of Antwerp and submitter of the GAMBL-AW system, provided a baseline of 62.4% using the same method (the 1.5% difference is most likely explained by how well the baseline systems dealt with multi-word constructions and hyphenated words).

4 Conclusion

As with the SENSEVAL-2 English all-words task, the supervised systems fared much better than

System	Precision/Recall
GAMBL-AW-S	.652
SenseLearner-S	.646
Koc University-S	.641
R2D2: English-all-words	.626
Meaning-allwords-S	.624
Meaning-simple-S	.610
upv-shmm-eaw-S	.609
LCCaw	.607
UJAEN-S	.590
IRST-DDD-00-U	.583
University of Sussex-Prob5	.572
University of Sussex-Prob4	.554
University of Sussex-Prob3	.551
DFA-Unsup-AW-U	.548
IRST-DDD-LSI-U	.501
KUNLP-Eng-All-U	.500
upv-unige-CIAOSENSO-eaw-U	.481
merl.system3	.458
upv-unige-CIAOSENSO2-eaw-U	.452
merl.system1	.450
IRST-DDD-09-U	.446
autoPS-U	.436
clr04-aw	.434
merl.system2	.359
autoPSNVs-U	.359
DLSI-UA-all-Nosu	.280

Table 1: "With U" scores; a -S or -U suffix after the system name indicates that the system was reported as supervised or unsupervised, respectively.

System	Precision	Recall
GAMBL-AW-S	.651	.651
SenseLearner-S	.651	.642
Koc University-S	.648	.639
R2D2: English-all-words	.626	.626
Meaning-allwords-S	.625	.623
Meaning-simple-S	.611	.610
LCCaw	.614	.606
upv-shmm-eaw-S	.616	.605
UJAEN-S	.601	.588
IRST-DDD-00-U	.583	.582
University of Sussex-Prob5	.585	.568
University of Sussex-Prob4	.575	.550
University of Sussex-Prob3	.573	.547
DFA-Unsup-AW-U	.557	.546
KUNLP-Eng-All-U	.510	.496
IRST-DDD-LSI-U	.661	.496
upv-unige-CIAOSENSO-eaw-U	.581	.480
merl.system3	.467	.456
upv-unige-CIAOSENSO2-eaw-U	.608	.451
merl.system1	.459	.447
IRST-DDD-09-U	.729	.441
autoPS-U	.490	.433
clr04-aw	.506	.431
autoPSNVs-U	.563	.354
merl.system2	.480	.352
DLSI-UA-all-Nosu	.343	.275

Table 2: "Without U" scores, sorted by recall; a -S or -U suffix after the system name indicates that the system was reported as supervised or unsupervised, respectively.

System Name	Email Contact
autoPS	dianam@sussex.ac.uk
autoPSNVs	dianam@sussex.ac.uk
clr04-aw	ken@clres.com
DFA-Unsup-AW	david@lsi.uned.es
DLSI-UA-Nosu	montoyo@dlsi.ua.es
GAMBL-AW	bart.decadt@ua.ac.be
IRST-DDD-00	strappa@itc.it
IRST-DDD-09	strappa@itc.it
IRST-DDD-LSI	strappa@itc.it
Koc University	dyuret@ku.edu.tr
KUNLP-Eng-All	hcseo@nlp.korea.ac.kr
LCCaw	parker@languagecomputer.com
Meaning	lluism@lsi.upc.es
Meaning simple	lluism@lsi.upc.es
merl.system1	bhiksha@merl.com
merl.system2	bhiksha@merl.com
merl.system3	bhiksha@merl.com
R2D2: EAW	montoyo@dlsi.ua.es
SenseLearner	rada@cs.unt.edu
UJAEN	mgarcia@ujaen.es
USussex-Prob3	Judita.Preiss@cl.cam.ac.uk
USussex-Prob4	Judita.Preiss@cl.cam.ac.uk
USussex-Prob5	Judita.Preiss@cl.cam.ac.uk
upv-shmm-eaw	amolina@dsic.upv.es
upv-CIAOSENSO	amolina@dsic.upv.es
upv-CIAOSENSO2	amolina@dsic.upv.es

Table 3: email contact for each system; sorted alphabetically.

the unsupervised systems (Palmer et al., 2001). In fact, all of the seven systems reported as supervised scored higher than any of the nine systems reported as unsupervised in both precision and recall (using either of the two scoring criteria).

The greatest difference between these results and those of the SENSEVAL-2 English all-words task is that a greater number of systems have now achieved scores at or above the baseline. While this result is encouraging, it seems that the best systems have a hit a wall in the 65-70% range. This is not surprising given the typical inter-annotator agreement of 70-75% for this task. We believe that further significant progress must await the development of resources with coarser-grained sense distinctions and with glosses that draw explicit contrasts between the senses – resources more suitable for the task at hand.

References

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Martha Palmer, Christiane Fellbaum, Scott Cotton, Lauren Delfs, and Hoa Trang Dang. 2001. English tasks: All-words and verb lexical sample. In *Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*, Toulouse, France, July.

Martha Palmer, Olga Babko-Malaya, and Hoa Trang Dang. 2004. Different granularities for different applications. In *Second Workshop on Scalable Natural Language Understanding Systems, HLT–NAACL*, Boston, MA, May.