English Tasks: All-Words and Verb Lexical Sample

Martha Palmer, Christiane Fellbaum, Scott Cotton,
Lauren Delfs, and **Hoa Trang Dang**
University of Pennsylvania
{mpalmer,fellbaum,cotton,lcdelfs,htd}@linc.cis.upenn.edu

Abstract

We describe our experience in preparing the lexicon and sense-tagged corpora used in the English all-words and lexical sample tasks of SENSEVAL-2.

1 Overview

The English lexical sample task is the result of a coordinated effort between the University of Pennsylvania, which provided training/test data for the verbs, and Adam Kilgarriff at Brighton, who provided the training/test data for the nouns and adjectives (see Kilgarriff, this issue). In addition, we provided the test data for the English all-words task. The pre-release version of WordNet 1.7 from Princeton was used as the sense inventory. Most of the revisions of sense definitions relevant to the English tasks were done prior to the bulk of the tagging.

The manual annotation for both the English all-words and verb lexical sample tasks was done by researchers and students in linguistics and computational linguistics at the University of Pennsylvania. All of the verbs in both the lexical sample and all-words tasks were annotated using a graphical tagging interface that allowed the annotators to tag instances by verb type and view the sentences surrounding the instances. Well over 1000 person hours went into the tagging tasks.

2 English All-Words Task

The test data for the English all-words task consisted of 5,000 words of running text from three Wall Street Journal articles representing varied domains from the Penn Treebank II. Annotators preparing the data were allowed to indicate at most one multi-word construction for each content word to be tagged, but could give multiple senses for the construction. In some cases, a multi-word construction was annotated with senses associated with just the head word of the phrase in addition to more specific senses based on the entire phrase. The annotations were done under a double-blind scheme by two linguistics students, and were then adjudicated and corrected by a different person.

Task participants were supplied with test data only, in the standard all-words format for SENSEVAL-2, as well as the original syntactic

System	Precision	Recall
SMUaw-	0.690	0.690
AVe-Antwerp	0.636	0.636
LIA-Sinequa-AllWords	0.618	0.618
david-fa-UNED-AW-T	0.575	0.569
david-fa-UNED-AW-U	0.556	0.550
gchao2-	0.475	0.454
gchao3-	0.474	0.453
Ken-Litkowski-clr-aw (*)	0.451	0.451
Ken-Litkowski-clr-aw	0.416	0.451
gchao-	0.500	0.449
cm.guo-usm-english-tagger2	0.360	0.360
magnini2-irst-eng-all	0.748	0.357
cmguo-usm-english-tagger	0.345	0.338
c.guo-usm-english-tagger3	0.336	0.336
agirre2-ehu-dlist-all	0.572	0.291
judita-	0.440	0.200
dianam-system3ospdana	0.545	0.169
dianam-system2ospd	0.566	0.169
dianam-system1	0.598	0.140
woody-IIT2	0.328	0.038
woody-IIT3	0.294	0.034
woody-IIT1	0.287	0.033

Table 1: System performance on English all-words task (fine-grained scores); (*) indicates system results that were submitted after the SENSEVAL-2 workshop and official deadline.

Christiane Fellbaum is at Princeton University, fellbaum@clarity.princeton.edu

and part-of-speech annotations from the Treebank. Table 1 shows the system performance on the task. Most of the systems tagged almost all the content words. This included not only indicating the appropriate sense from the WordNet 1.7 pre-release (as it stood at the time of annotation), but also marking multi-word constructions appropriate to the corresponding sense tags. If given a perfect lemmatizer, a simple baseline strategy which does not attempt to find the satellite words in multi-word constructions, but which simply tags each head word with the first WordNet sense for the corresponding Treebank part-of-speech tag, would result in precision and recall of about 0.57.

3 English Lexical Sample Task

The data for the verb lexical sample task came primarily from the Penn Treebank II Wall Street Journal corpus. However, where that did not supply enough samples to approximate 75+15*n instances per verb, where n is the number of senses for the verb, we supplemented with British National Corpus instances. We did not find sentences for every sense of every word we tagged. We also sometimes found sentences for which none of the available senses were appropriate, and these were discarded. The instances for each verb were partitioned into training/test data using a ratio of 2:1.

We also grouped the nouns, adjectives and verbs for the lexical sample task, attempting to be explicit about the criteria for each grouping. In particular, the criteria for grouping verbs included differences in semantic classes of arguments, differences in the number and type of arguments, whether an argument refers to a created entity or a resultant state, whether an event involves concrete or abstract entities or constitutes a mental act, whether there is a specialized subject domain, etc. All of the verbs were grouped by two or more people, with differences being reconciled. In some cases the groupings of the verbs are identical to the existing WordNet groupings; in some cases they are quite different. The nouns and adjectives were grouped by the primary annotator in the project; WordNet does not have comparable groups for nouns and adjectives.

These groupings were used for coarse-grained scoring, under the framework of SENSEVAL-1. After the SENSEVAL-2 workshop, participants were invited to retrain their systems on the groups; only a handful of participants chose to do this, and in the end the results were uniformly only slightly better than training on the fine-grained senses with coarse-grained scoring.

Table 2 shows the system performance on just the verbs of the lexical sample task. For comparison we ran several simple baseline algorithms that had been used in SENSEVAL-1, including RANDOM, COMMONEST, LESK, LESK-DEFINITION, and LESK-CORPUS (Kilgarriff and Rosenzweig, 2000). In contrast to SENSEVAL-1, in which none of the competing systems performed significantly better than the highest baseline (LESK-CORPUS), the best-performing systems this time performed well above the highest baseline.

Overall, the performance of the systems was much lower than in SENSEVAL-1. Several factors may have contributed to this. In addition to the use of fine-grained WordNet senses instead of the smaller Hector sense inventory from SENSEVAL-1, most of the verbs included in this task were chosen specifically because we expected them to be difficult to tag. There was also generally less training data made available to the systems (ignoring outliers, there were on average twice as many training samples for each verb in SENSEVAL-1 as there were in SENSEVAL-2). Table 3 shows the correspondence between test data size (half of training data size), entropy, and system performance for each verb.

4 Annotating the Gold Standard

The annotators made every effort to match the target word to a WordNet sense both syntactically and semantically, but sometimes this could not be done. Given a conflict between syntax and semantics, the annotators opted to match semantics. For example, the word "train" has an intransitive sense ("undergo training or instruction in preparation for a particular role, function, or profession") as well as a related (causative) transitive sense ("create by training and teaching"). Instances of "train" that were interpreted as having a dropped object were tagged with the transitive sense even though the overt syntax did not match the sense definition.

Some sentences seemed to fit equally well with two different senses, often because of am-

System	P	R
agirre3-ehu-dlist-best	0.846	0.229
magnini-irst-eng-sample	0.660	0.138
kunlp-	0.576	0.576
jhu-english-JHU-final (*)	0.566	0.566
SMUls-	0.563	0.563
LIA-Sinequa-Lexsample	0.535	0.535
manning-cs224n	0.523	0.523
agirre3-ehu-dlist-all	0.514	0.493
talp-TALP	0.513	0.513
umcp-englishl-	0.494	0.493
jhu-english-JHU-ENGLISH	0.489	0.489
montoyo-Univ.-Alicante-System	0.486	0.480
jhu-english-JHU	0.485	0.485
tdp1-duluth3	0.465	0.465
tdp1a-duluthC	0.453	0.453
tdp1-duluth5	0.450	0.450
tdp1-duluth4	0.446	0.446
baseline-lesk-corpus	0.445	0.445
tdp1-duluth2	0.440	0.440
tdp1a-duluthA	0.439	0.439
tdp1-duluth1	0.437	0.437
tdp1a-duluthB	0.404	0.404
baseline-commonest	0.403	0.403
david-fal-UNED-LS-T	0.388	0.387
david-fal-UNED-LS-U	0.288	0.287
Haynes-IIT2	0.233	0.232
Haynes-IIT1	0.220	0.220
Kenneth-Litkowski-clr-ls	0.218	0.218
Haynes-IIT2 (*)	0.199	0.192
Haynes-IIT1 (*)	0.193	0.186
baseline-lesk	0.181	0.181
michael-oakes.suss2	0.094	0.094
baseline-lesk-def	0.088	0.088
baseline-random	0.085	0.085

Table 2: System precision (P) and recall (R) for English verb lexical sample task (fine-grained scores); (*) indicates system results that were submitted after the SENSEVAL-2 workshop and official deadline.

biguous context; others did not fit well under any sense. One of the solutions employed in these cases was the assignment of multiple sense tags. The taggers would choose two senses (on rare occasions, even three) that they felt made an approximation of the correct sense when used in combination. Sometimes this strategy was also used in arbitration, when it was decided that neither tagger's tag was better than the other. The taggers tried to use this strategy sparingly and chose single tags whenever possible.

Often, a particular verb yielded multiple in-

Verb	Size	Entropy	Fine	Coarse
ferret	1	0.00	0.913	0.913
collaborate	30	0.44	0.898	0.898
wander	50	0.96	0.619	0.786
face	93	1.09	0.690	0.785
replace	45	1.62	0.471	0.860
use	76	1.68	0.558	0.682
begin	280	1.76	0.625	0.625
treat	44	2.10	0.453	0.543
live	67	2.35	0.455	0.476
match	42	2.35	0.398	0.620
train	63	2.60	0.394	0.492
drift	32	2.77	0.327	0.354
dress	59	2.89	0.434	0.679
serve	51	3.02	0.404	0.445
drive	42	3.03	0.308	0.528
leave	66	3.06	0.317	0.428
develop	69	3.17	0.301	0.456
see	69	3.28	0.278	0.317
wash	12	3.31	0.343	0.535
work	60	3.54	0.303	0.442
keep	67	3.62	0.336	0.353
call	66	3.68	0.246	0.457
play	66	3.80	0.323	0.345
find	68	3.81	0.178	0.285
carry	66	3.97	0.279	0.332
strike	54	4.06	0.248	0.331
pull	60	4.24	0.255	0.414
draw	41	4.60	0.195	0.264
turn	67	4.79	0.216	0.327

Table 3: Test corpus size, entropy (base 2) of tagged data, and average system recall for each verb, using fine-grained and coarse-grained scoring.

stances of what was clearly a salient sense, but one not found in WordNet. One of the results was that sentences that should have received a clear sense tag ended up with something rather ad hoc, and often inconsistent. One of the most notorious examples was "call," which had no sense that fit sentences like "The restaurant is called Marrakesh." WordNet contains some senses related to this one. One sense refers to the bestowing of a name; another to informal designations; another to greetings and vocatives. But there is no sense in WordNet for simply stating something's name without additional connotations, and the gap possibly caused some inconsistencies in the annotation. All these senses belonged to the same group, and if the annotators had been allowed to tag with the more general group sense, there may

have been less inconsistency.

It has been well-established that sense-tagging is a very difficult task (Kilgarriff, 1997; Hanks, 2000), even for experienced human taggers. If the sense inventory has gaps or redundancies, or if some of the sense glosses have ambiguous wordings, choosing the correct sense can be all but impossible. Even if the annotator is working with a very good entry, unforeseen instances of the word always arise.

The degree of polysemy does not affect the relative difficulty of tagging, at least not in the way it is often thought. Very polysemous words, such as "drive," are not necessarily harder to tag than less polysemous words like "replace." The difficulty of tagging depends much more on other aspects of the entry and of the word itself. Often very polysemous words *are* quite difficult to tag, because they are more likely to be underspecified or occur in novel uses; however, "replace," with four senses, proved a difficult verb to tag, while "play," with thirty-five senses, was relatively straightforward.

In many ways, the grouped senses are very helpful for the sense-tagger. Grouping similar senses allows the sense-tagger to study side-by-side the senses that are perhaps most likely to be confused, which is helpful when the differences between the senses are very subtle. However, it would be a poor idea to attempt to tag a corpus using *only* the groups, and not the finer sense distinctions, because often some of the senses included in a group will have some properties that the others do not; it is always better to make the finest distinction possible and not just assign the same tag to everything that seems close.

Inter-annotator agreement figures for the human taggers are quite low. However, in some respects they are not quite as low as they seem. Some of the apparent discrepancies were simply the result of a technical error: the annotator accidentally picked the wrong tag, perhaps choosing one of its neighbors. Other differences resulted from the sense inventories themselves. Sometimes the taggers interpreted the wording of a given sense definition in different ways, which caused them to choose different tags, but does not entail that they had interpreted the instances differently; in fact, discussion of such cases usually revealed that the taggers had interpreted the instances themselves in the sam way. Additional apparent discrepancies resulte from the various strategies for dealing with case in which there was no single proper sense i WordNet. This was the case when an instanc in the corpus was underspecified so as to al low multiple appropriate interpretations. Thi resulted in (a) multiple tags by one or bot taggers, and (b) each tagger making a differ ent choice. Here, again, the taggers often hac the same interpretation of the instance itself but because the sense inventory was insufficien for their needs, they were forced to find differen strategies. Sometimes, in fact, one tagger would double-tag a particular instance while the second tagger chose a single sense that matched one of the two selected by the first annotator. This is considered a discrepancy for statistical purposes, but clearly reflects similar interpretations on the part of the annotators.

In the most recent evaluation, with two new annotators tagging against the Gold Standard, the best fine-grained agreement figures for verbs were in the 70's, similar to Semcor figures. However, when we used the groupings to do a more coarse-grained evaluation, and counted a match between a single tag and a member of a double tag as correct, the human annotator agreement figures rose to 90%.

5 Acknowledgments

Support for this work was provided by the National Science Foundation (grants NSF-9800658 and NSF-9910603), DARPA (grant 535626), and the CIA (contract number 2000*SO53100*000). We would also like to thank Joseph Rosenzweig for building the annotation tools, and Susanne Wolff for contribution to the manual annotation.

References

Patrick Hanks. 2000. Do word meanings exist? *Computers and the Humanities*, 34(1-2), April. Special Issue on SENSEVAL.

A. Kilgarriff and J. Rosenzweig. 2000. Framework and results for English SENSEVAL. *Computers and the Humanities*, 34(1-2), April. Special Issue on SENSEVAL.

Adam Kilgarriff. 1997. I don't believe in word senses. *Computers and the Humanities*, 31(2).