English Lexical Sample Task Description

Adam Kilgarriff
ITRI, University of Brighton
Brighton, UK
adam@itri.bton.ac.uk

The English lexical sample task (adjectives and nouns) for SENSEVAL 2 was set up according to the same principles as for SENSEVAL-1, as reported in (Kilgarriff and Rosenzweig, 2000). (Adjectives and nouns only, because the data preparation for the verbs lexical sample was undertaken alongside that for the English all-words task, and is reported in Palmer et al (this volume). All discussion below up to the Results section covers only adjectives and nouns.)

1 Lexical sample

The lexicon was sampled to give a range of low, medium and high frequency words (see Table 1). These were all different words to the ones used in SENSEVAL 1.

2 Corpus choice

For the most part, the British National Corpus (New edition) was used. (The new edition has the advantage that it is available worldwide, so all participants had the opportunity of obtaining it for system training.) Our goal was to match this source, containing British English, with another, of American English. In the event, only limited quantities of corpus data for American English were available without copyright complications, so the lion's share of the data was from the BNC with a limited quantity from the Wall Street Journal.

In accordance with standard SENSEVAL procedure, the goal was to have $75 + 15n + 6m$ instances for each lexical-sample word, where n is the number of senses the word has and m is the number of multiword expressions that the word is part of (both, of course, relative to a specific lexicon). In practice numbers varied slightly, as instances were deleted because they had the wrong part of speech or were otherwise unusable. See Table 1 for actual numbers of senses, multiwords expressions and instances.

3 Lexicon choice

Here lay the biggest contrast with the SENSEVAL-1 task, which had used Oxford University Press's experimental HECTOR lexicon. This time, in response to popular acclaim, WordNet was used.

Since SENSEVAL was first mooted, in 1997, WordNet-or-not-WordNet has been a recurring theme. In favour was the argument that it was already very widely used, almost a *de facto* standard. The argument against concerned its sense distinctions. WordNet, like thesauruses but unlike standard dictionaries, is organised around groups of words of similar meanings (*synsets*), not around words (with their various meanings). This means that the priority for the lexicographer is building coherent synsets rather than the coherent analysis of the various meanings of a particular word. The writer of a thesaurus does not need to pay as much attention to the distinction between two senses of a word, as the writer of a dictionary. Word sense disambiguation is a task which needs clear and well-motivated sense distinctions. In English SENSEVAL-1, WordNet was not used because of concerns that it did not provide clean enough sense distinctions.

While HECTOR provided good sense distinctions, it was unsatisfactory in that it did not cover the whole lexicon so there was no possibility of scaling up. The case for WordNet – that it was already integrated into so much NLP and WSD work – still stood, so the decision was made to use WordNet. To guard against cases where WordNet made a distinction between two meanings, but it was not clear what the distinction was, all the words in the lexical sample had their entries reviewed by a

Word	Ss	Mwe	inst	ITA
ADJS: lexical sample size: 15				
blind	3	21	163	89.6
colorless	2	0	103	94.2
cool	6	1	158	92.1
faithful	3	0	70	94.6
fine	9	6	212	84.0
fit	3	0	86	85.0
free	8	36	247	79.2
graceful	2	0	85	72.6
green	7	80	284	86.6
local	3	12	113	89.1
natural	10	37	309	72.4
oblique	2	5	86	96.4
simple	7	19	196	67.8
solemn	2	0	77	84.1
vital	4	7	112	93.7
ALL ADJS			2301	83.4
NOUNS: lexical sample size: 29				
art	5	35	294	78.5
authority	7	6	276	84.3
bar	13	57	455	87.3
bum	4	0	137	91.7
chair	4	35	207	92.8
channel	7	10	218	84.8
child	4	16	193	92.3
church	3	21	192	88.0
circuit	6	31	255	93.5
day	9	82	434	76.3
detention	2	5	95	98.7
dyke	2	0	86	96.5
facility	5	9	172	89.5
fatigue	4	6	128	97.7
feeling	6	5	153	77.0
grip	7	3	153	85.2
hearth	3	1	96	85.0
holiday	2	9	93	90.5
lady	3	27	158	74.1
material	5	39	209	85.1
mouth	8	10	179	88.7
nation	3	10	112	90.5
nature	5	8	138	86.7
post	8	33	236	87.7
restraint	6	3	136	80.4
sense	5	37	160	87.1
spade	3	7	98	95.1
stress	5	7	118	74.7
yew	2	15	85	97.1
ALL NOUNS			5266	86.3
ALL			7567	85.5

Table 1: Lexical sample: rubric for column headers: Ss=number of fine-grained senses; Mwe = number of multi-word expressions which the word participates in (as *bear* participates in WordNet headword *polar bear*); inst = number of instances tagged; ITA = inter-tagger agreement (fine-grained).

lexicographer, with a view particularly to merging insufficiently-distinct senses. It was initially unclear how these revisions would relate to the publicly available version of WordNet (at that time, WordNet 1.6). We are very grateful to the Princeton WordNet team (George Miller, Christiane Fellbaum and Randee Tengi) for their help at this point; they agreed to incorporate our proposed revisions into a new version of WordNet (1.7) which was then made available in time (despite some very tight deadlines) for the SENSEVAL competition.

WordNet 1.7 was not available as a complete object at the time of the gold standard production, in Spring 2001, but the entries for the lexical sample words were fixed at that point. For each lexical sample entry, we produced an HTML version for the lexicographers to work from. In addition to all the relevant information in WordNet, this had a mnemonic for each sense, so that taggers could use mnemonics when doing the tagging, rather than easily-forgotten, easily-confused sense numbers. The mnemonics were selected by a lexicographer.

4 Gold standard production

Once the corpus sources and lexical entries were fixed, work could proceed with the Gold-Standard tagging.[1]

First, a team of three professional lexicographers and fourteen students and others was recruited. Recruitment proceeded as follows: an aptitude test was set up on the web. The test involved sense-tagging some corpus instances (taken from SENSEVAL-1, so the gold-standard answers were known). Email postings were made asking interested people to visit the website and take the test. All applicants scoring sufficiently well on the test were then offered work, on a piecework basis.

An HTML version of the corpus for a word was prepared. This comprised a series of ten-sentence stretches of text, with one word in the last of the sentences highlighted; that was the word to be sense-tagged. The files were HTML versions of the XML files used for test and training data.

A tagger was emailed the lexical entry and corpus for a word. They then tagged it, and

[1]The tagging was supported by a grant from EPSRC, the UK funding council, under GR/R02337/01 (MATS).

returned, by email, a file of answers. These files were checked automatically, and if they contained 'answers' which were not possible answers for the word, the suspect items were automatically emailed back to the tagger for correction.

The tagger guidelines are available along with other resources for the English-lexical-sample task. They developed in the course of the exercise; when a tagger asked a pertinent questions, I circulated the question and my answer to all taggers and incorporated them into the guidelines.

As in SENSEVAL-1, "Unassignable" and "Proper-name" tags were always available alongside regular tags, and taggers were told to put down more than one tag, where multiple tags were equally applicable. Taggers were also asked to mark items where the part of speech was wrong; these were then deleted from the dataset.

5 Tagger agreement procedures and scores

As in all exercises where a gold standard corpus is the goal, it was necessary to have all data tagged by more than one person. The question then arises, how many taggings does each item need? The algorithm adopted here was:

1. send item out to two taggers
2. if they agree completely, **stop; return agreed answer**
3. else, send out to another tagger
4. is there one or more tag that two agree on?
5. if yes, **stop; return all tags which two people agree on**
6. if no, return to step 3

Thus, in simple cases, a minimum of effort was used, but in difficult cases, more opinions were obtained. The number of taggings per items is shown below. Note that the algorithm stops at step 2 if both taggers agree on one tag, or if both taggers agree on two or more tags.

Taggings	Number	%
2	5032	66.5
3	2446	32.3
4	86	1.1
5	4	0.05

3 taggers' answers			GS	cases
A	A	B	A	651
A	A;B	A	A	550
A	A;B	B	A;B	209
A	A;B	A;B	A;B	189
A	A;B	C	A	162
A	A	A;B;C	A	67
A	A;B	A;C	A	51
A	A	B;C	A	44
A;B	A;C	C	A;C	41
A;B	A:B;C	C	A;B;C	38

Table 2: Patterns of (dis)agreement for 3-tagger cases. GS = gold standard tagging arising from these human taggings. ";" used as separator where a tagger (or the gold standard) gave multiple tags.

Of the 5032 two-tagger items, in 4688 cases, the taggers agreed on one tag; in 340 cases, on two tags; and in 4 cases, on three tags.

For the 2446 cases which were tagged three times, 136 were cases where all three taggers agreed perfectly (so, had the algorithm been followed to the letter, the item would not have been tagged a third time; such cases were caused by delays in taggers returning answers.) The common patterns amongst the remainder are shown in Table 2.

For the 86 cases with four taggers, half the cases were {A, A, B, C} taggings.

Fine-grained inter-tagger agreement (ITA) figures was calculated using the same scoring algorithm as for the systems.[2] For each pair of taggers tagging an instance, two scores were calculated, one with the one answer as the key, the other with the other. For each instance, scores were normalised so that the maximum score for each corpus instance was one, however many times it had been tagged. The overall ITA was 85.5%. A breakdown by word and by word class is given in Table 1.[3]

[2] All ITA figures and other results reported in this paper refer to fine-grained sense distinctions. The grouping of senses into coarse-grained categories took place independently of the gold-standard preparation, which was based entirely on fine sense distinctions.

[3] Kappa was not calculated because there were various ways in which it might have been calculated, so it was unclear which was appropriate, and it would have introduced more complication than clarification. Also

As argued in (Kilgarriff and Rosenzweig, 2000) (also (Kilgarriff, 1999)) the inter-tagger agreement figure for a gold standard is of less interest than the replicability figure: if a completely different team of taggers used the same methodology to do the same task, what would the agreement level between the two teams' outputs be? It is the replicability figure, rather than ITA, which defines an upper bound for the task. We have not yet had time to conduct such a study.

6 Task organisation

The organisation followed standard SENSEVAL procedure. The data was prepared in XML using SENSEVAL DTDs, with the data for each word split in a ration of 2:1 between training and test data. Data distribution, results uploads, baselines and scoring were handled at UPenn (see paper by Cotton and Edmonds).

7 Results

Results are presented in the table below. Owing to space constraints, where a team submitted multiple systems with similar results, only the best result is shown. Full results are available at the SENSEVAL website, as are decodings of system names. At the SENSEVAL workshop (5–6 July 2001) it was agreed that there should also be a later deadline (end July 2001) so that 'egregious bugs' could be fixed. In order to honour both standard practice in evaluation exercises (eg, no extension of deadlines) and also the agreement made at the workshop, both results sets are presented, with later-deadline results marked with (R) as a suffix to the name.

There has not yet been time for an analysis of the results. The one comment that does seem pertinent is the contrast with the English-lexical-sample task in SENSEVAL-1. The tasks were organised in similar ways, and some of the systems were improved versions of systems participating in 1998. Yet the performance of the best systems has, apparently, dropped around 14%. We may well ask, why?

We believe the drop is due to the choice of lexicon. As discussed above, using WordNet for SENSEVAL has drawbacks. High-accuracy word sense disambiguation is only possible where the lexicon makes clear and well-motivated sense distinctions, and provides sufficient information about the distinctions for the disambiguation algorithm to build on. An implication for future WSD research is that it is time to turn our attention from algorithms, to sense distinctions.

PR	ATT	System
Supervised systems		
.82	28	BCU ehu-dlist-best
.67	25	IRST
.64	100	JHU (R)
.64	100	SMUls
.63	100	KUNLP
.62	100	Stanford-CS224
.61	100	Sinequa-LIA SCT
.59	100	TALP
.57	98	BCU ehu-dlist-all
.57	100	Duluth-3
.57	100	UMD-SST
.50	100	UNED LS-T
.42	98	Alicante
Supervised baselines		
.51	100	Base Lesk
.48	100	Base Commonest
Unsupervised systems		
.58	55	ITRI-WASPS
40	100	UNED-LS-U
.29	100	CLresearch DIMAP
.25	99	IIT-2 (R)
Unsupervised baselines		
.16	100	Base Lesk-defs
.14	100	Base random

Table 3: PR=system precision; ATT= percentage of cases for which an answer was returned ("attempted").

the figures shown, unlike kappa figures, have the merit of being directly comparable with system performance scores.

References

Adam Kilgarriff and Joseph Rosenzweig. 2000. Framework and results for English SENSEVAL. *Computers and the Humanities*, 34(1–2):15–48. Special Issue on SENSEVAL, edited by Adam Kilgarriff and Martha Palmer.

Adam Kilgarriff. 1999. 95% replicability for manual word sense tagging. In *Proc. EACL*, pages 277–278, Bergen, June.