SemEval-2007 Task 10: English Lexical Substitution Task

Diana McCarthy
University of Sussex
Falmer, East Sussex
BN1 9QH, UK
dianam@sussex.ac.uk

Roberto Navigli
University of Rome "La Sapienza"
Via Salaria, 113
00198 Roma, Italy
navigli@di.uniroma1.it

Abstract

In this paper we describe the English Lexical Substitution task for SemEval. In the task, annotators and systems find an alternative substitute word or phrase for a target word in context. The task involves both finding the synonyms and disambiguating the context. Participating systems are free to use any lexical resource. There is a subtask which requires identifying cases where the word is functioning as part of a multiword in the sentence and detecting what that multiword is.

1 Introduction

Word sense disambiguation (WSD) has been described as a task in need of an application. Whilst researchers believe that it will ultimately prove useful for applications which need some degree of semantic interpretation, the jury is still out on this point. One problem is that WSD systems have been tested on fine-grained inventories, rendering the task harder than it need be for many applications (Ide and Wilks, 2006). Another significant problem is that there is no clear choice of inventory for any given task (other than the use of a parallel corpus for a specific language pair for a machine translation application).

The lexical substitution task follows on from some previous ideas (McCarthy, 2002) to examine the capabilities of WSD systems built by researchers on a task which has potential for NLP applications. Finding alternative words that can occur in given contexts would potentially be useful to many applications such as question answering, summarisation, paraphrase acquisition (Dagan et al., 2006), text simplification and lexical acquisition (McCarthy, 2002). Crucially this task does not specify the inventory for use beforehand to avoid bias to one predefined inventory and makes it easier for those using automatically acquired resources to enter the arena. Indeed, since the systems in SemEval did not know the candidate substitutes for a word before hand, the lexical resource is evaluated as much as the context based disambiguation component.

2 Task set up

The task involves a lexical sample of nouns, verbs, adjectives and adverbs. Both annotators and systems select one or more substitutes for the target word in the context of a sentence. The data was selected from the English Internet Corpus of English produced by Sharoff (2006) from the Internet (http://corpus.leeds.ac.uk/internet.html). This is a balanced corpus similar in flavour to the BNC, though with less bias to British English, obtained by sampling data from the web. Annotators are not provided with the PoS (noun, verb, adjective or adverb) but the systems are. Annotators can provide up to three substitutes but all should be equally as good. They are instructed that they can provide a phrase if they can't think of a good single word substitute. They can also use a slightly more general word if that is close in meaning. There is a "NAME" response if the target is part of a proper name and "NIL" response if annotators cannot think of a good substitute. The subjects are also asked to identify

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 48–53,
Prague, June 2007. ©2007 Association for Computational Linguistics

if they feel the target word is an integral part of a phrase, and what that phrase was. This option was envisaged for evaluation of multiword detection. Annotators did sometimes use it for paraphrasing a phrase with another phrase. However, for an item to be considered a constituent of a multiword, a majority of at least 2 annotators had to identify the same multiword.[1]

The annotators were 5 native English speakers from the UK. They each annotated the entire dataset. All annotations were semi-automatically lemmatised (substitutes and identified multiwords) unless the lemmatised version would change the meaning of the substitute or if it was not obvious what the canonical version of the multiword should be.

2.1 Data Selection

The data set comprises 2010 sentences, 201 target words each with 10 sentences. We released 300 for the trial data and kept the remaining 1710 for the test release. 298 of the trial, and 1696 of the test release remained after filtering items with less than 2 non NIL and non NAME responses and a few with erroneous PoS tags. The words included were selected either manually (70 words) from examination of a variety of lexical resources and corpora or automatically (131) using information in these lexical resources. Words were selected from those having a number of different meanings, each with at least one synonym. Since typically the distribution of meanings of a word is strongly skewed (Kilgarriff, 2004), for the test set we randomly selected 20 words in each PoS for which we manually selected the sentences [2] (we refer to these words as MAN) whilst for the remaining words (RAND) the sentences were selected randomly.

2.2 Inter Annotator Agreement

Since we have sets of substitutes for each item and annotator, pairwise agreement was calculated between each pair of sets ($p1, p2 \in P$) from each possible pairing (P) as $\frac{\sum_{p1,p2 \in P} \frac{p_1 \cap p_2}{p_1 \cup p_2}}{|P|}$

Pairwise inter-annotator agreement was 27.75%. 73.93% had modes, and pairwise agreement with the mode was 50.67%. Agreement is increased if we remove one annotator who typically gave 2 or 3 substitutes for each item, which increased coverage but reduced agreement. Without this annotator, inter-annotator agreement was 31.13% and 64.7% with mode.

Multiword detection pairwise agreement was 92.30% and agreement on the identification of the exact form of the actual multiword was 44.13%.

3 Scoring

We have 3 separate subtasks 1) **best** 2) **oot** and 3) **mw** which we describe below. [3] In the equations and results tables that follow we use P for precision, R for recall, and $Mode\ P$ and $Mode\ R$ where we calculate precision and recall against the substitute chosen by the majority of annotators, provided that there is a majority.

Let H be the set of annotators, T be the set of test items with 2 or more responses (non NIL or NAME) and h_i be the set of responses for an item $i \in T$ for annotator $h \in H$.

For each $i \in T$ we calculate the mode (m_i) i.e. the most frequent response provided that there is a response more frequent than the others. The set of items where there is such a mode is referred to as TM. Let A (and AM) be the set of items from T (or TM) where the system provides at least one substitute. Let $a_i : i \in A$ (or $a_i : i \in AM$) be the set of guesses from the system for item i. For each i we calculate the multiset union (H_i) for all h_i for all $h \in H$ and for each unique type (res) in H_i will have an associated frequency ($freq_{res}$) for the number of times it appears in H_i.

For example: Given an item (id 9999) for *happy;a* supposing the annotators had supplied answers as follows:

annotator	responses
1	glad merry
2	glad
3	cheerful glad
4	merry
5	jovial

[1] Full instructions given to the annotators are posted at http://www.informatics.susx.ac.uk/research/nlp/mccarthy/files/instructions.pdf.

[2] There were only 19 verbs due to an error in automatic selection of one of the verbs picked for manual selection of sentences.

[3] The scoring measures are as described in the document at http://nlp.cs.swarthmore.edu/semeval/tasks/task10/task10documentation.pdf released with our trial data.

then H_i would be *glad glad glad merry merry cheerful jovial*. The res with associated frequencies would be *glad 3 merry 2 cheerful 1* and *jovial 1*.

best measures This requires the **best** file produced by the system which gives as many guesses as the system believes are fitting, but where the credit for each correct guess is divided by the number of guesses. The first guess in the list is taken as the best guess (bg).

$$P = \frac{\sum_{a_i:i \in A} \frac{\frac{\sum_{res \in a_i} freq_{res}}{|a_i|}}{|H_i|}}{|A|} \quad (1)$$

$$R = \frac{\sum_{a_i:i \in T} \frac{\frac{\sum_{res \in a_i} freq_{res}}{|a_i|}}{|H_i|}}{|T|} \quad (2)$$

$$Mode\ P = \frac{\sum_{bg_i \in AM} 1\ if\ bg = m_i}{|AM|} \quad (3)$$

$$Mode\ R = \frac{\sum_{bg_i \in TM} 1\ if\ bg = m_i}{|TM|} \quad (4)$$

A system is permitted to provide more than one response, just as the annotators were. They can do this if they are not sure which response is better, however systems will maximise the score if they guess the most frequent response from the annotators. For P and R the credit is divided by the number of guesses that a system makes to prevent a system simply hedging its bets by providing many responses. The credit is also divided by the number of responses from annotators. This gives higher scores to items with less variation. We want to emphasise test items with better agreement.

Using the example for *happy;a* id 9999 above, if the system's responses for this item was *glad; cheerful* the credit for a_{9999} in the numerator of P and R would be $\frac{\frac{3+1}{2}}{7} = .286$

For $Mode\ P$ and $Mode\ R$ we use the system's first guess and compare this to the mode of the annotators responses on items where there was a response more frequent than the others.

oot measures This allows a system to make up to 10 guesses. The credit for each correct guess is not divided by the number of guesses. This allows for the fact that there is a lot of variation for the task and we only have 5 annotators. With 10 guesses there is a better chance that the systems find the responses of these 5 annotators. There is no ordering of the guesses and the $Mode$ scores give credit where the mode was found in one of the system's 10 guesses.

$$P = \frac{\sum_{a_i:i \in A} \frac{\sum_{res \in a_i} freq_{res}}{|H_i|}}{|A|} \quad (5)$$

$$R = \frac{\sum_{a_i:i \in T} \frac{\sum_{res \in a_i} freq_{res}}{|H_i|}}{|T|} \quad (6)$$

$$Mode\ P = \frac{\sum_{a_i:i \in AM} 1\ if\ any\ guess \in a_i = m_i}{|AM|} \quad (7)$$

$$Mode\ R = \frac{\sum_{a_i:i \in TM} 1\ if\ any\ guess \in a_i = m_i}{|TM|} \quad (8)$$

mw measures For this measure, a system must identify items where the target is part of a multiword and what the multiword is. The annotators do not all have linguistics background, they are simply asked if the target is an integral part of a phrase, and if so what the phrase is. Sometimes this option is used by the subjects for paraphrasing a phrase of the sentence, but typically it is used when there is a multiword. For scoring, a multiword item is one with a majority vote for the same multiword with more than 1 annotator identifying the multiword.

Let MW be the subset of T for which there is such a multiword response from a majority of at least 2 annotators. Let $mw_i \in MW$ be the multiword identified by majority vote for item i. Let $MWsys$ be the subset of T for which there is a multiword response from the system and $mwsys_i$ be a multiword specified by the system for item i.

$$detection\ P = \frac{\sum_{mwsys_i \in MWsys} 1\ if\ mw_i\ exists\ at\ i}{|MWsys|} \quad (9)$$

$$detection\ R = \frac{\sum_{mwsys_i \in MW} 1\ if\ mw_i\ exists\ at\ i}{|MW|} \quad (10)$$

$$identification\ P = \frac{\sum_{mwsys_i \in MWsys} 1\ if\ mwsys_i = mw_i}{|MWsys|} \quad (11)$$

$identification\ R =$

$$\frac{\sum_{mwsys_i \in MW} 1\ if\ mwsys_i = mw_i}{|MW|} \quad (12)$$

3.1 Baselines

We produced baselines using WordNet 2.1 (Miller et al., 1993a) and a number of distributional similarity measures. For the WordNet **best** baseline we found the best ranked synonym using the criteria 1 to 4 below in order. For WordNet **oot** we found up to 10 synonyms using criteria 1 to 4 in order until 10 were found:

1. Synonyms from the first synset of the target word, and ranked with frequency data obtained from the BNC (Leech, 1992).
2. synonyms from the hypernyms (verbs and nouns) or closely related classes (adjectives) of that first synset, ranked with the frequency data.
3. Synonyms from all synsets of the target word, and ranked using the BNC frequency data.
4. synonyms from the hypernyms (verbs and nouns) or closely related classes (adjectives) of all synsets of the target, ranked with the BNC frequency data.

We also produced **best** and **oot** baselines using the distributional similarity measures l1, jaccard, cosine, lin (Lin, 1998) and αSD (Lee, 1999) [4]. We took the word with the largest similarity (or smallest distance for αSD and l1) for **best** and the top 10 for **oot**.

For **mw** detection and identification we used WordNet to detect if a multiword in WordNet which includes the target word occurs within a window of 2 words before and 2 words after the target word.

4 Systems

9 teams registered and 8 participated, and two of these teams (SWAG and IRST) each entered two systems, we distinguish the first and second systems with a 1 and 2 suffix respectively.

The systems all used 1 or more predefined inventories. Most used web queries (HIT, MELB, UNT) or web data (Brants and Franz, 2006) (IRST2, KU, SWAG1, SWAG2, USYD, UNT) to obtain counts for disambiguation, with some using algorithms to derive domain (IRST1) or co-occurrence (TOR) information from the BNC. Most systems did not use sense tagged data for disambiguation though MELB did use SemCor (Miller et al., 1993b) for filtering infrequent synonyms and UNT used a semi-supervised word sense disambiguation combined with a host of other techniques, including machine translation engines.

5 Results

In tables 1 to 3 we have ordered systems according to R on the **best** task, and in tables 4 to 6 according to R on **oot**. We show all scores as percentages i.e. we multiply the scores in section 3 by 100. In tables 3 and 6 we show results using the subset of items which were i) NOT identified as multiwords (NMWT) ii) scored only on non multiword substitutes from both annotators and systems (i.e. no spaces) (NMWS). Unfortunately we do not have space to show the analysis for the MAN and RAND subsets here. Please refer to the task website for these results. [5] We retain the same ordering for the further analysis tables when we look at subsets of the data. Although there are further differences in the systems which would warrant reranking on an individual analysis, since we combined the subanalyses in one table we keep the order as for 1 and 4 respectively for ease of comparison.

There is some variation in rank order of the systems depending on which measures are used. [6] KU is highest ranking on R for **best**. UNT is best at finding the mode, particularly on **oot**, though it is the most complicated system exploiting a great many knowledge sources and components. IRST2 does well at finding the mode in **best**. The IRST2 **best** R score is lower because it supplied many answers for each item however it achieves the best R score on the **oot** task. The baselines are outperformed by most systems. The WordNet baseline outperforms those derived from distributional methods. The distributional methods, especially lin, show promising results given that these methods are automatic and

[4] We used 0.99 as the parameter for α for this measure.

[5] The task website is at http://www.informatics.sussex.ac.uk/research/nlp/mccarthy/task10index.html.

[6] There is not a big difference between P and R because systems typically supplied answers for most items.

Systems	P	R	$Mode\ P$	$Mode\ R$
KU	12.90	12.90	20.65	20.65
UNT	12.77	12.77	20.73	20.73
MELB	12.68	12.68	20.41	20.41
HIT	11.35	11.35	18.86	18.86
USYD	11.23	10.88	18.22	17.64
IRST1	8.06	8.06	13.09	13.09
IRST2	6.95	6.94	20.33	20.33
TOR	2.98	2.98	4.72	4.72

Table 1: **best** results

Systems	P	R	$Mode\ P$	$Mode\ R$
WordNet	9.95	9.95	15.28	15.28
lin	8.84	8.53	14.69	14.23
l1	8.11	7.82	13.35	12.93
lee	6.99	6.74	11.34	10.98
jaccard	6.84	6.60	11.17	10.81
cos	5.07	4.89	7.64	7.40

Table 2: **best** baseline results

don't require hand-crafted inventories. As yet we haven't combined the baselines with disambiguation methods.

Only HIT attempted the **mw** task. It outperforms all baselines from WordNet.

5.1 Post Hoc Analysis

Choosing a lexical substitute for a given word is not clear cut and there is inherently variation in the task. Since it is quite likely that there will be synonyms that the five annotators do not think of we conducted a post hoc analysis to see if the synonyms selected by the original annotators were better, on the whole, than those in the systems responses. We randomly selected 100 sentences from the subset of items which had more than 2 single word substitutes, no NAME responses, and where the target word was

	NMWT		NMWS	
Systems	P	R	P	R
KU	13.39	13.39	14.33	13.98
UNT	13.46	13.46	13.79	13.79
MELB	13.35	13.35	14.19	13.82
HIT	11.97	11.97	12.55	12.38
USYD	11.68	11.34	12.48	12.10
IRST1	8.44	8.44	8.98	8.92
IRST2	7.25	7.24	7.67	7.66
TOR	3.22	3.22	3.32	3.32

Table 3: Further analysis for **best**

Systems	P	R	$Mode\ P$	$Mode\ R$
IRST2	69.03	68.90	58.54	58.54
UNT	49.19	49.19	66.26	66.26
KU	46.15	46.15	61.30	61.30
IRST1	41.23	41.20	55.28	55.28
USYD	36.07	34.96	43.66	42.28
SWAG2	37.80	34.66	50.18	46.02
HIT	33.88	33.88	46.91	46.91
SWAG1	35.53	32.83	47.41	43.82
TOR	11.19	11.19	14.63	14.63

Table 4: **oot** results

Systems	P	R	$Mode\ P$	$Mode\ R$
WordNet	29.70	29.35	40.57	40.57
lin	27.70	26.72	40.47	39.19
l1	24.09	23.23	36.10	34.96
lee	20.09	19.38	29.81	28.86
jaccard	18.23	17.58	26.87	26.02
cos	14.07	13.58	20.82	20.16

Table 5: **oot** baseline results

	NMWT		NMWS	
Systems	P	R	P	R
IRST2	72.04	71.90	76.19	76.06
UNT	51.13	51.13	54.01	54.01
KU	48.43	48.43	49.72	49.72
IRST1	43.11	43.08	45.13	45.11
USYD	37.26	36.17	40.13	38.89
SWAG2	39.95	36.51	40.97	37.75
HIT	35.60	35.60	36.63	36.63
SWAG1	37.49	34.64	38.36	35.67
TOR	11.77	11.77	12.22	12.22

Table 6: Further analysis for **oot**

	HIT		WordNet BL	
	P	R	P	R
detection	45.34	56.15	43.64	36.92
identification	41.61	51.54	40.00	33.85

Table 7: MW results

	good	reasonable	bad
sys	9.07	19.08	71.85
origA	37.36	41.01	21.63

Table 8: post hoc results

not one of those identified as a multiword (i.e. a majority vote by 2 or more annotators for the same multiword as described in section 2). We then mixed the substitutes from the human annotators with those of the systems. Three fresh annotators[7] were given the test sentence and asked to categorise the randomly ordered substitutes as good, reasonable or bad. We take the majority verdict for each substitute, but if there is one reasonable and one good verdict, then we categorise the substitute as reasonable. The percentage of substitutes for systems (sys) and original annotators (origA) categorised as good, reasonable and bad by the post hoc annotators are shown in table 8. We see the substitutes from the humans have a higher proportion of good or reasonable responses by the post hoc annotators compared to the substitutes from the systems.

6 Conclusions and Future Directions

We think this task is an interesting one in which to evaluate automatic approaches of capturing lexical meaning. There is an inherent variation in the task because several substitutes may be possible for a given context. This makes the task hard and scoring is less straightforward than a task which has fixed choices. On the other hand, we believe the task taps into human understanding of word meaning and we hope that computers that perform well on this task will have potential in NLP applications. Since a pre-defined inventory is not used, the task allows us to compare lexical resources as well as disambiguation techniques without a bias to any predefined inventory. It is possible for those interested in disambiguation to focus on this, rather than the choice of substitutes, by using the union of responses from the annotators in future experiments.

7 Acknowledgements

We acknowledge support from the Royal Society UK for funding the annotation for the project, and for a Dorothy Hodgkin Fellowship to the first author. We also acknowledge support to the second author from INTEROP NoE (508011, 6^{th} EU FP). We thank the annotators for their hard work. We thank Serge Sharoff for the use of his Internet corpus, Julie Weeds for the software we used for producing the distributional similarity baselines and Suzanne Stevenson for suggesting the **oot** task .

[7] Again, these were native English speakers from the UK.

References

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram corpus version 1.1. Technical Report.

Ido Dagan, Oren Glickman, Alfio Gliozzo, Efrat Marmorshtein, and Carlo Strapparava. 2006. Direct word sense matching for lexical substitution. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, July. Association for Computational Linguistics.

Nancy Ide and Yorick Wilks. 2006. Making sense about sense. In Eneko Agirre and Phil Edmonds, editors, *Word Sense Disambiguation, Algorithms and Applications*, pages 47–73. Springer.

Adam Kilgarriff. 2004. How dominant is the commonest sense of a word? In *Proceedings of Text, Speech, Dialogue*, Brno, Czech Republic.

Lillian Lee. 1999. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 25–32.

Geoffrey Leech. 1992. 100 million words of English: the British National Corpus. *Language Research*, 28(1):1–13.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, Madison, WI.

Diana McCarthy. 2002. Lexical substitution as a task for wsd evaluation. In *Proceedings of the ACL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 109–115, Philadelphia, USA.

George Miller, Richard Beckwith, Christine Fellbaum, David Gross, and Katherine Miller, 1993a. *Introduction to WordNet: an On-Line Lexical Database*. ftp://clarity.princeton.edu/pub/WordNet/5papers.ps.

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

Serge Sharoff. 2006. Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, 11(4):435–462.