Lang Resources & Evaluation (2009) 43:139–159
DOI 10.1007/s10579-009-9084-1

The English lexical substitution task

Diana McCarthy · Roberto Navigli

Published online: 26 February 2009

Abstract Since the inception of the Senseval series there has been a great deal of debate in the word sense disambiguation (WSD) community on what the right sense distinctions are for evaluation, with the consensus of opinion being that the distinctions should be relevant to the intended application. A solution to the above issue is lexical substitution, i.e. the replacement of a target word in context with a suitable alternative substitute. In this paper, we describe the English lexical substitution task and report an exhaustive evaluation of the systems participating in the task organized at SemEval-2007. The aim of this task is to provide an evaluation where the sense inventory is not predefined and where performance on the task would bode well for applications. The task not only reflects WSD capabilities, but also can be used to compare lexical resources, whether man-made or automatically created, and has the potential to benefit several natural-language applications.

Keywords Lexical substitution · Word sense disambiguation · SemEval-2007

1 Introduction

The English lexical substitution task was conceived as a means of examining the issue of word sense representation by giving participants a free reign over the lexical inventories used on a task that evaluates the inventories and also contextual disambiguation. The identification of the meaning of a word in context, namely word sense disambiguation (WSD), is a task with a long history in the

D. McCarthy (✉)
The University of Sussex, Falmer, East Sussex, UK
e-mail: dianam@sussex.ac.uk

R. Navigli
The University of Rome "La Sapienza", Rome, Italy
e-mail: navigli@di.uniroma1.it

computational linguistics field (Ide and Véronis 1998). While it has been argued that word senses might boost applications such as information retrieval (Schütze and Pederson 1995; Stokoe et al. 2003) and machine translation (Chan et al. 2007; Carpuat and Wu 2007) there has been a long standing debate as to the appropriateness of predefined sense inventories for computational purposes (Palmer 2000; Hanks 2000; Ide and Wilks 2006; Kilgarriff 2006). Many are concerned with granularity because fine-grained distinctions are not easy to make by either humans or machines (Navigli et al. 2007; Navigli 2006; Palmer et al. 2007; Ide and Wilks 2006) although note that there are also indications that some more subtle distinctions might be useful for some applications (Stokoe 2005). Generally what is agreed is that the distinctions should befit the application. An outstanding problem is determining the appropriateness of an inventory for a given application, other than machine translation where the sense distinctions can be determined by the target languages (Chan et al., 2007; Carpuat and Wu 2007; Resnik and Yarowsky 2000).

While it is very encouraging that we are at last seeing some results that indicate the benefits of WSD for machine translation (Chan et al. 2007; Carpuat and Wu 2007), there are many monolingual applications that also stand to gain from being able to determine the meaning of a word in context. It is not appropriate to use parallel data for making generalisations about sense distinctions for monolingual understanding because the sense distinctions made by translations depend entirely on which language pairs are used; ultimately any distinction that can be made will be made by some pair of languages (Palmer et al. 2007).

The lexical substitution task was proposed for SemEval-2007 following earlier ideas on the topic (McCarthy 2002). The task requires systems to produce a substitute word for a word in context. For example a substitute of *game* might be given for the word *match* in the following sentence:

After the match, replace any remaining fluid deficit to prevent problems of chronic dehydration throughout the tournament.

Crucially the systems are not given the possible substitutes for a target word in advance. This means that the inventory is as much part of the evaluation as the disambiguation component. It also means that systems can select the inventory of their choice or induce a new one and we can investigate the merits of different inventories.

There are a great many applications that either require or might benefit from systems to find a replacement word or phrase in context, for example summarisation or text simplification. Capabilities at the lexical substitution task would also be useful for systems that recognise paraphrases, for example, for question answering and textual entailment (Dagan et al. 2005).

The structure of this paper is as follows. In Sect. 2 we describe the task and how it was set up, with motivations for the choices that were made for the data sampling, annotation, sub-tasks and scoring. We finish the section with a description of the baselines provided for the task. In Sect. 3 we describe the participating systems with a detailed breakdown of the resources and techniques adopted by the systems. In Sect. 4 we provide an overview of the results and analysis by part-of-speech (PoS). We also provide analysis to contrast the ability of systems to find substitutes compared with the ability to determine the right substitute for the right context. We follow this with a section on post hoc analysis where we sought the judgements of fresh annotators after the SemEval-2007 evaluation period to quantify how much the systems were able to come up with good substitutes that the original annotators did not think of.

2 The task

The lexical substitution task involves annotators and systems finding a substitute for a word in the context of a sentence. The motivation for our lexical substitution task was to allow a representation of meaning that does not necessitate a pre-defined listing of senses and where systems are free to select an appropriate representation for the task. By not prescribing the inventory, we avoid the need for mappings between inventories which can mask results and we provide a level playing field for systems which use a non-standard inventory such as those which automatically induce meanings (Schütze 1998; Pantel and Lin 2002). Lexical substitution involves both:

1. finding the set of candidate substitutes for the word
2. finding the best candidate given the context of the word token

There are a large number of pre-defined (Fellbaum 1998; Roget 1911; Barnard 1986) and automatically induced (Pantel and Lin 2002) inventories of word meaning which provide information on semantically related words. This task also provides a means of comparing performance of these different resources for finding good candidates.

2.1 The data

The data for this task was collected for a sample 201 words: nouns, verbs, adjectives and adverbs. For each word we collected ten sentences containing that word in the given PoS. The data was selected from the English Internet Corpus (EIC) produced by Sharoff (2006) from the web (http://corpus.leeds.ac.uk/internet.html). This is a balanced corpus similar in flavour to the British National Corpus (BNC) (Leech 1992), though with less bias to British English, obtained by sampling data from the web. We selected EIC on the grounds that it was freely available and because it would reduce any bias due to systems over-training on this data, since currently most systems use the BNC (Leech 1992), Gigaword (Graff 2003) or Google Web 1T (Brants and Franz 2006). We removed noisy data (e.g. lists of items) automatically by selecting for sentences between 6 and 79 words long and where the target word was not capitalised (or all caps) when not at the start of a sentence. We used the PoS tags supplied with the corpus to obtain sentences for the word and PoS combination and then manually screened the sentences for problems.

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.[1] 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 provided that they

[1] Available from http://www.informatics.sussex.ac.uk/research/groups/nlp/mccarthy/task10index.html.

have more than one meaning 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 chose the sentences[2] whilst for the remaining words the sentences were selected randomly.

2.2 Annotation

The annotators were five native English speakers from the UK; three women and two men. Three had a linguistics background and two did not. They were given full instructions[3] before annotating the entire data set in their own time on a web interface. They could log in and out at any point. Annotators were instructed to find a word as a replacement for the target word in context that preserves the meaning of the original word as much as possible. The interface displayed one sentence at a time with the target word highlighted in bold. The annotators were led through all the sentences for one word before the next and all received the sentences in the same order. The annotators could return to any sentence at any time and change their response (just as the systems can). Annotators were not provided with the PoS of the target (noun, verb, adjective or adverb) though the systems were.

The annotators were allowed to provide up to three substitutes however, they were instructed that all substitutes given for an item should be equally valid. They were instructed that they could provide a phrase if they could not think of a good single word substitute and also that they could use a slightly more general word if that is close in meaning. There was a "NAME" response if the target is part of a proper name and a "NIL" response where the annotator could not think of a decent substitute. Items were only used for scoring where there were at least two non "NAME" and non "NIL" responses for the item from the set of annotators. Annotators were permitted to consult a dictionary or thesaurus of their choosing, and if they did they were also asked to tick a "used dictionary" check box which held its value for that word. Since a given sentence may include an occurrence of a target word within a multiword (phrase functioning as a word), the annotators were instructed to identify sentences where they felt the target word is an integral part of a phrase, and what that phrase was. This data was used for multiword subtasks which are not described here due to space limitations, see McCarthy and Navigli (2007) and McCarthy (2008) for further details.

2.2.1 An example

In Fig. 1, we provide an example of the substitutes given by annotators for the noun *charge*. The figure shows the substitutes provided by the annotators for each of the 10 sentences in the test data. Thus for the first sentence for *charge*:

In the event of a chemical spill, 3/4's of the children know that they should evacuate (leave area) as advised on radio, TV, or by people in <u>*charge*</u>.

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

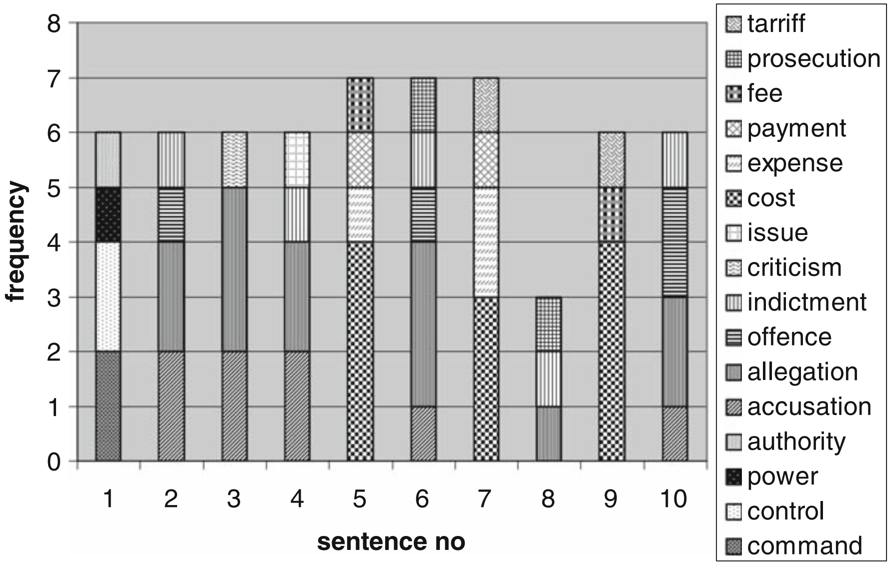

Fig. 1 Annotator substitutes provided for charge (n)

the substitutes selected by the five annotators were:

Annotator	1	2	3	4	5
substitutes	*control, command*	*control*	*authority*	*power*	*command*

Note that the frequency of substitutes can exceed the number of annotators (five) since each annotator can provide up to three responses if they feel all fit the sentence equally well. The variation of substitutes highlights the different meanings of the words, for example *charge* has meanings **power**, **allegation** and **cost** in the ten sentences.

2.2.2 Inter-annotator agreement

We calculated inter-annotator agreement on the test set using 2 measures:

1. pairwise agreement between annotators (PA)
2. pairwise agreement with the mode, i.e. the most frequent response (PA_m)

Let H be the set of annotators, T be the set of test items with two or more responses (non NIL or NAME) and h_i be the set of responses for an item $i \in T$ for annotator $h \in H$. This set is empty if the response from the annotator was “NAME” or “NIL”. Let P_i be the set of all possible pairwise combinations of the sets of non-empty responses from any pair of annotators in H for item i. Let $\{h_i, h'_i\} \in P_i$ be one combination of annotators responses for a pair of annotators h, h'. So pairwise agreement between annotators is calculated as:

Table 1 Inter-annotator agreement by PoS

PoS	#	*PA*	% With modes	PA_m
Noun	497	28.4	74.4	52.2
Verb	440	25.2	72.3	48.6
Adjective	468	24.0	72.7	47.4
Adverb	298	36.4	77.5	56.1
All	1703	27.7	73.9	50.7

$$PA = \sum_{i \in T} \frac{\sum_{\{h_i, h'_i\} \in P_i} \frac{h_i \cap h'_i}{h_i \cup h'_i}}{|P_i| \cdot |T|} \qquad (1)$$

Thus, pairwise agreement measures the average proportion of all the paired responses (in P_i) for which the two paired annotators gave the same response.

For each $i \in T$ we calculate the mode (m_i) 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 T_m. Pairwise agreement with the mode is calculated as:

$$PA_m = \frac{\sum_{i \in T_m} \sum_{h_i\,:\,h \in H} \frac{1 \text{ if } m_i \in h_i}{|h_i|}}{|HT_m|} \qquad (2)$$

where HT_m is the set of all non empty h_i for $h \in H$ for the items in T_m.

Pairwise inter-annotator agreement (*PA*) was 27.75%. 73.93% of the test items had modes, and PA_m was 50.67%. Agreement is increased if we remove one annotator who typically gave two or three substitutes for each item, which increased coverage but reduced agreement. Without this annotator, inter-annotator agreement was 31.13%, and 64.7% with the mode. Agreement is low compared to a task with a fixed inventory and reflects the fact that there is no clear right and wrong answer for many items, but several possibilities where some may be better than others, reflected in the popularity of the choice among annotators. Table 1 shows the agreement figures by PoS, with the number of items in T (#) for each PoS and the percentage of items with a mode (4th column). Note that there was most agreement for adverbs, then nouns, verbs and the lowest agreement was for the adjectives, presumably because there is typically a larger variety of potential substitutes for adjectives.

2.3 The test data and gold standard

From the 300 items released as trial data, there were 298 items in *T*, that is having at least two substitutes from the set of annotators. From the 1710 items released for the test set, 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.

All substitutes were semi-automatically lemmatised unless the lemmatised version would change the meaning of the substitute.

2.4 Scoring

There are two separate subtasks which will be described below.[4]

best the system gives as many guesses as it believes are fitting, but where the credit for each correct guess is divided by the number of guesses. The very best should be given first.

oot (out of ten) up to 10 guesses with no penalising for multiple guesses to cope with fact that we only have five annotators and systems may come up with a larger, but equally valid, set of substitutes

In the equations and results tables that follow we use R for recall, and *Mode R* where we calculate recall against the substitute chosen by the majority of annotators, provided that there is a majority.[5]

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

As explained in subsection 2.2.2, 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 T_m. Let A (and A_m) be the set of items from T (or T_m) where the system provides at least one substitute. Let $a_i : i \in A$ (or $a_i : i \in A_m$) 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	1	2	3	4	5
substitutes	*glad, merry*	*glad*	*cheerful, glad*	*merry*	*jovia*

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.*

2.4.1 **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*).

[4] In the SemEval-2007 task, there was also a third subtask on multiwords. Only one system participated in the multiword subtask, so we do not to describe it here. The scoring measures for all three subtasks are as described in the document at http://nlp.cs.swarthmore.edu/semeval/tasks/task10/task10documentation.pdf released with our trial data.

[5] We also calculated precision over the items attempted by a system which can be contrasted with recall which includes all items. Since systems typically left out only a few items we do not report the precision figures here due to lack of space as the results are similar.

We calculate recall (R) as the average annotator response frequency of substitutes found by the system over all items in T:

$$R = \frac{\sum_{a_i:i \in T} \frac{\sum_{res \in a_i} freq_{res}}{|a_i| \cdot |H_i|}}{|T|} \tag{3}$$

Mode recall (*Mode R*) is calculated as the proportion of matches between the best guess and the mode from the annotator's responses over the number of items with a mode ($|T_m|$):

$$Mode\ R = \frac{\sum_{bg_i \in T_m} 1\ if\ bg_i = m_i}{|T_m|} \tag{4}$$

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 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 R would be $\frac{3+1}{2 \cdot 7} = .286$

For *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.

2.4.2 **oot** Measures

Out-of-ten measures (**oot**) allow a system to make up to ten guesses. The credit for each correct guess is not divided by the number of guesses. This accounts for the fact that there is a lot of variation for the task and we only have five annotators. With ten guesses there is a better chance that the systems find the responses of these five annotators. There is no ordering of the guesses and the *Mode* scores give credit where the mode was found as any one of the system's ten guesses. Our measures for **oot** are:

$$R = \frac{\sum_{a_i:i \in T} \frac{\sum_{res \in a_i} freq_{res}}{|H_i|}}{|T|} \tag{5}$$

$$Mode\ R = \frac{\sum_{a_i:i \in TM} 1 \text{ if } m_i \in a_i}{|T_m|} \tag{6}$$

The above formulas are similar to Eqs. (3) and (4) (in contrast to the latter, the number of guesses $|a_i|$ for an item i, and the position of m_i in a_i do not influence the formulas for **oot**).

2.5 Baselines

We produced baselines using WordNet 2.1 (Fellbaum 1998), the Roget's thesaurus (Roget 1911) and a number of distributional similarity measures.

For the WordNet **best** baseline we found the best ranked synonym[6] using the criteria 1 to 4 below in order:[7]

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 BNC frequency data.
3. Synonyms from all synsets of the target word, and ranked with 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.

For the Roget baseline, the **best** ranking substitute was chosen among all possible synonyms of the target word ranked with the BNC frequency data (notice that in Roget's thesaurus entries are not ranked).

The **best** baselines produced using distributional similarity measures were obtained by applying the measures to the contexts provided by grammatical relation data obtained from parsing the written portion of the BNC using the RASP system (Briscoe and Carroll 2002). Different grammatical relations were used for obtaining synonyms for target words with different PoS. For nouns we used verbs for which the nouns occurred as direct objects or subjects and adjective or noun modifiers. For verbs, we used the nouns acting as direct objects or subjects. For adjectives we used the modified noun or modifying adverb and for adverbs, we used the modified adjective or verb. We used the l1, jaccard, cosine, lin (Lin 1998) and αSD[8] (Lee 1999) distributional similarity measures. We took the word with the largest similarity (or smallest distance for αSD and l1) for **best**.

3 Systems

There were eight participating teams. The teams all participated in both tasks, expect that MELB only participated in **best** and SWAG2 only participated in **oot**.

We use the name of the team when referring to the system, however two teams (SWAG and IRST) each entered two systems. In the descriptions and results that follow, we distinguish the first and second systems of these teams with a 1 and 2 suffix respectively. In this section we provide an overview of the systems to facilitate

[6] We only used single words as substitutes for the baseline as we did not have frequency data for multiwords.

[7] For WordNet **oot** we found up to ten synonyms using the same criteria in order until ten were found. We do not report the **oot** baselines here due to lack of space and because we observed a similar pattern to the **best** baseline.

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

Table 2 Sources for candidate substitutes

System	WordNet	Macquarie	Roget	Other
MELB (Martinez et al. 2007)	√			
HIT (Zhao et al. 2007)	√			
UNT (Hassan et al. 2007)	√			Encarta
IRST1 (Giuliano et al. 2007)	√			OAWT
IRST2 (Giuliano et al. 2007)	√			OAWT
KU (Yuret 2007)			√	
SWAG1 (Dahl et al. 2007)			√	
SWAG2 (Dahl et al. 2007)			√	
USYD (Hawker 2007)	√	√		Web 1T
TOR (Mohammad et al. 2007)		√		

comparisons between systems when interpreting the results. For full details of the systems please use the reference to the SemEval-2007 proceedings given in Table 2.

The systems all used one or more predefined lexical inventories for obtaining candidate substitutes. The inventories that were used by each team are indicated in Table 2. The lexical inventories used are WordNet (Fellbaum 1998),[9] Roget's New Millenium Thesaurus (Thesaurus.com 2007), the Macquarie thesaurus (Barnard 1986), the Oxford American Writer's thesaurus (OAWT)(Lindberg 2004) and the Microsoft Encarta Thesaurus (an online encyclopedia and thesaurus). USYD was the only system to supplement candidates from predefined resources (WordNet and Macquarie) with candidates from corpus data (the Web 1T (Brants and Franz 2006)), though the candidates from the predefined thesaurus were selected in preference to those from the corpus data.

Table 3 lists the corpus data used by each system. Most used search engine queries (HIT, MELB, UNT) or the Web 1T corpus (IRST2, KU, SWAG1, SWAG2, USYD, UNT) to obtain counts for disambiguation. The BNC was also used for deriving morphological information (UNT), for domain modelling using latent semantic analysis (LSA) (IRST1), for creating frequency vectors as representations for the senses in Roget (SWAG2) and for pointwise mutual information (*pMI*) to measure association between Macquarie categories and the target context and also frequency to rank the candidates within a category (TOR). Only one system (UNT) used sense-tagged data for WSD in a semi-supervised system (Mihalcea and Csomai 2005) which used SemCor (Miller et al. 1993) for training. One other system MELB used the frequency information from SemCor. In this case the labelled data was used for filtering infrequent synonyms from the list of candidates.

We will not attempt to duplicate the more detailed descriptions of the systems provided in the SemEval-2007 proceedings and referenced here in Table 2. Clearly as well as the actual resources and approaches taken, there were other differences in the methods, for example any smoothing taken for the n-gram approach. We will highlight similarities and differences that seem pertinent when analysing the results

[9] USYD and HIT used version 2.1, the others based on WordNet all used 2.0.

Table 3 Resources used for disambiguation

System	Search engine	Web 1T	BNC	SemCor	Other
MELB	n-gram			filter	
HIT	n-gram				
UNT	n-gram	n-gram	morph	WSD	TE + Wiki
IRST1			LSA		
IRST2		n-gram			
KU		n-gram			
SWAG1		n-gram			
SWAG2		n-gram	freq vectors		
USYD		*pMI*			
TOR			*pMI* + freq		

in the following section. Note that UNT was a complicated system involving a combination of modules which used two thesauri, two n-gram models a semi-supervised WSD system described in (Mihalcea and Csomai 2005), a Wikipedia LSA model (Wiki) and Google and Altavista translation engines (TE) to translate the test sentence back and forth between English and French.

4 Results

In this section we provide results of the submitted systems along with new analysis. In the first subsection 4.1 we give results using the official scorer for this task and described above in Sect. 2.4. In addition to the results provided in (McCarthy and Navigli 2007) we provide PoS analysis for both **best** and **oot** subtasks. We do not show the breakdown of systems for performance on items without a multiword target or substitute. In (McCarthy and Navigli 2007), we show that all systems did better when the target word was not identified by the annotators as being a multiword, and did better on all items when only considering the substitutes that were not multiwords.

The lexical substitution task has two main components: finding the candidate substitutes and selecting the one that best fits the target context. In two subsections, we examine the capabilities of the systems for both (i) finding substitutes for the target word and PoS and (ii) matching the substitutes to the sentence. In subsection 4.2 we compare the set of substitute words over all sentences for the target word and PoS combination from the gold-standard and the system *bg* (first choice i.e. best guess) responses. We only use the *bg* response from the **best** subtask to simplify the analysis. We examine the extent that the systems find the full set of substitute types from the annotators within the *bg* responses that they provide for the ten sentences for the word and PoS combination. We examine this in the context of the total number of *bg* substitute types that they provide for a word and PoS combination, and the extent that they find substitute types as *bg* that the annotators also find which are not found by the other systems. In subsection 4.3 we look at the disambiguation performance of systems where they have found the gold-standard mode within their set of *bg* responses for ten sentences for the word and PoS. That is, where the system has access to the best answer for the word and PoS and the issue is whether or not they can select the best substitute for the context.

4.1 Results using the task scorer

In this section we show the results of all submitted systems on the **best** and **oot** tasks in two respective subsections. We only show recall, because most systems provided a response for all the items so the precision was only marginally better than recall. The exceptions to this were the USYD and particularly SWAG2 and SWAG1 which did not provide answers for all items and precision was a few percentage points better than recall. For each system, we provide separate scores for each PoS. We show all scores as percentages i.e. we multiply the scores in Sect. 2.4 by 100.

Firstly, it is important to stress that these were the results of the officially submitted systems. There were several apparent errors in the system responses which could easily be corrected. For example, the responses from USYD included non-word substitutes such as punctuation and digits which could be filtered as unlikely substitutes. TOR had many repetitions of the target word as the substitute, again these errors could have been removed with filtering. The IRST2 system performed poorly on **best** because it supplied many answers to each item so credit was divided. The first answer was typically a good guess as shown by the better performance of IRST2 on the mode scores.

4.1.1 Results for the **best** *task*

The figures for **best** using recall and mode recall are given in Tables 4 and 5 respectively. In Table 6 we analyse the performance of all baselines on the **best** subtask for the recall measure to see the relative performance of the baselines by PoS. The last column displays the order of performance for the different PoS: nouns (n), verbs (v), adjectives (a) and adverbs (r).

We see from the results of **best** recall and mode recall in Tables 4 and 5 that systems that rely on WordNet seem to do very well on adverbs. This follows from the high performance of the WordNet baseline on this PoS shown below in Table 6.

The **best** recall score (see Eq. 3) divides the credit for each answer by the sum of the responses of the human annotators ($|H_i|$) to allow more credit to be given to the items with less variation, i.e. where there is most agreement. This means that the recall score for the **best** task depends on the frequency of the most frequent substitute from the annotators[10] and the total frequency of all responses from the annotators. The theoretical upper bound for this score is calculated as:

$$best\ upper\ bound = \frac{\sum_{i \in T} \frac{freq_{most\ freq\ substitute_i}}{|H_i|}}{|T|} = 0.4576 \qquad (7)$$

[10] In the case of more than one substitute ranked highest by frequency, the recall score is limited by the frequency of any substitute that shares the highest rank.

Table 4 Results for **best** recall

Systems	All	Nouns	Verbs	Adjectives	Adverbs	Order
KU	12.90	12.14	10.68	13.92	15.85	ranv
UNT	12.77	12.26	7.90	12.25	21.63	rnav
MELB	12.68	9.41	9.01	12.94	23.09	ranv
HIT	11.35	11.91	6.47	9.54	20.43	rnav
USYD	10.88	11.01	8.31	9.60	16.46	rnav
IRST1	8.06	8.29	6.20	7.81	10.81	rnav
IRST2	6.94	5.77	4.65	6.89	12.33	ranv
TOR	2.98	2.79	0.99	4.04	4.59	ranv
WordNet bl	9.95	8.14	7.16	6.99	21.69	rnav
Lin bl	8.53	12.52	5.16	7.97	7.76	narv

Table 5 Results for **best** mode

Systems	All	Nouns	Verbs	Adjectives	Adverbs	Order
KU	20.65	18.82	18.79	21.41	24.89	ranv
UNT	20.73	19.94	11.46	19.88	35.62	rnav
MELB	20.41	13.76	14.33	22.02	36.48	ravn
HIT	18.86	19.66	11.15	16.51	31.33	rnav
USYD	17.64	17.98	13.69	15.29	25.75	rnav
IRST1	13.09	13.20	11.15	10.70	18.88	rnva
IRST2	20.33	19.10	14.33	21.41	28.76	ranv
TOR	4.72	3.93	0.96	7.95	6.44	arnv
WordNet BL	15.28	11.80	10.51	11.62	32.19	rnav
Lin BL	14.23	21.35	7.96	14.37	11.59	narv

Table 6 Baseline results for **best** recall

Systems	All	Nouns	Verbs	Adjectives	Adverbs	Order
WordNet	9.95	8.14	7.16	6.99	21.69	rnva
Lin	8.53	12.52	5.16	7.97	7.76	narv
l1	7.82	10.22	6.14	7.32	7.13	narv
αSD	6.74	9.39	2.99	8.50	5.15	narv
jaccard	6.60	8.86	4.37	5.96	7.15	nrav
cos	4.89	6.79	1.99	5.14	5.62	nrav
Roget	4.65	1.99	5.47	4.85	7.51	rvan

That is, we concur with Yuret (2007) on the upper bound for the **best** measure (45.76%), though note of course that this upper bound is theoretical and assumes a human could find the most frequent substitute selected by all five annotators. The upper bound will undoubtedly be lower because of human variability on this task, see the PA_m scores in Table 1. For the mode scores[11] the theoretical upper bound is 100% because credit is not divided by $|H_i|$, see Eq. (4). We discuss the **oot** upper bound in the section below.

Table 6 shows how the baselines described in Sect. 2.5 perform by PoS. While the WordNet baseline outperforms those obtained using automatically acquired thesauri, it is interesting to note that although WordNet outperforms the automatic thesauri on verbs and adverbs, the Lin baseline is better on nouns and the αSD thesaurus is better on adjectives. Indeed three out of the five automatically induced resources outperform WordNet on adjectives and nouns. Furthermore the Lin automatically induced thesaurus obtained the best results on nouns even compared to all the participating systems, which all attempt disambiguation in contrast to this rudimentary baseline. This is probably due to the fact that rarer senses, which confuse systems more than they help, are not well represented in the nearest neighbours of distributional thesauri. The l1 thesaurus is the automatic thesaurus that comes closest to the WordNet performance on verbs. WordNet does better on verbs than adjectives, contrary to the results of the systems using it for candidates, this is probably due to the impact of ambiguity. Roget meanwhile outperformed many of the distributional baselines for verbs and adverbs, but did poorly on adjectives and nouns in particular. This is undoubtedly because Roget offers a vast amount of choice, particularly for nouns and adjectives, and since it does not have a sense order (as WordNet does) it performs poorly without some form of contextual disambiguation.

4.1.2 *Results for the* ***oot*** *task*

The results for **oot** using recall and mode recall are shown in Tables 7 and 8. It was noted when analysing the results for **oot**, that some systems (IRST2, USYD and TOR) had included duplicates in the ten responses provided for some of the items. We had not anticipated this when designing the task and scorer. The effect of duplicates on the **oot** scores are that it is possible to get inflated scores which might then exceed 100% because the credit for each of the human answers ($freq_{res}$) is used for each of the duplicates, so a system that provides duplicates for responses that the annotators have provided will do better. For mode scores, providing duplicates is more likely to reduce a system's score because credit for each item simply requires a match between the mode from the annotators and any of the ten responses. Options for a match are reduced with duplicates.[12] While we could simply remove duplicates from the system responses, this would depart from the official scoring and would leave systems which did supply duplicates at a disadvantage. To avoid direct comparison with those that do not duplicate any responses, we partition the systems accordingly in the results reported here.

Duplicates in the **oot** task affect the theoretical upper bound. If one does not permit duplicates in the ten responses for any item then the theoretical upper bound for the **oot**

[11] Recall that these are only calculated on items where there is a mode.

[12] To highlight the problem of duplicates we have added a warning in the release version of the scorer which indicates where a duplicate is found and states that systems that include duplicates should NOT be compared with those that do not on **oot**.

Table 7 **oot** Recall

Systems	All	Nouns	Verbs	Adjectives	Adverbs	Order	IwD
IRST2	68.90	57.66	46.49	68.90	120.66	ranv	1232
USYD	34.96	33.14	41.10	29.96	36.71	vrna	443
TOR	11.19	9.94	6.12	10.21	22.28	ranv	371
UNT	49.19	48.07	44.24	47.80	60.54	rnav	0
KU	46.15	40.84	39.78	51.07	56.72	ranv	0
IRST1	41.20	38.48	32.18	43.12	56.07	ranv	0
SWAG2	34.66	22.63	31.56	42.19	47.46	ravn	0
HIT	33.88	32.13	29.25	29.22	50.89	rnva	0
SWAG1	32.83	27.95	28.75	42.19	32.33	arvn	0

Table 8 **oot** Mode recall

Systems	All	Nouns	Verbs	Adjectives	Adverbs	Order	IwD
IRST2	58.54	56.46	45.22	61.16	75.97	ranv	1232
USYD	42.28	40.17	37.26	42.51	51.93	ranv	443
TOR	14.63	12.64	5.10	15.90	28.76	ranv	371
UNT	66.26	64.04	59.24	66.36	78.97	ranv	0
KU	61.30	55.34	52.55	67.28	73.82	ranv	0
IRST1	55.28	52.81	43.63	56.57	72.96	ranv	0
HIT	46.91	44.66	39.81	40.37	69.10	rnav	0
SWAG2	46.02	30.06	41.72	55.35	63.09	ravn	0
SWAG1	43.82	37.64	38.54	55.35	44.21	arvn	0

task[13] is 100% because the entire set H_i can potentially be found within 10 responses allowed by this task since there is no item where $|H_i| > 10$. If one allows for duplicates then this increases the theoretical upper bound for this task since if one used the most frequent word type in H_i for all ten answers one would obtain ten times the **best** upper bound equation (7). That is the theoretical upper bound is 457.6%. Note however that the systems that have included duplicates may have done so unwittingly, and they do not do so for all items. In Tables 7 and 8 we have given the results of IRST1, USYD, TOR before the others as these systems include duplicates. The number of items with duplicate substitutes in the ten responses is shown in the column IwD.

4.2 Overlap analysis

The lexical substitution task involves both finding the substitutes and disambiguation of the target in context to find an appropriate substitute given the context. This

[13] We have not tried to calculate human agreement on the **oot** task because the gold-standard is the same as **best** and it is not realistic for humans to come up with ten substitutes for a given item. The **oot** task was envisaged as a way of compensating for the fact that we only have five annotators and there could be more substitutes than they can think of so it allows systems a better chance of finding the substitutes provided by the five annotators.

subsection analyses how the systems did at finding the substitute types across all sentences using the union of all responses for the word and PoS given by the annotators i.e. the distinct word types in H_i.[14] For this ‘overlap’ analysis we only looked at items where the target word was not considered to be a multiword and at single word substitutes (from both systems and humans). For each word and PoS combination (*wp*) we found the union of all substitute types from the annotators over the ten sentences for that word and PoS (GU_{wp}) and the union from the system for the word and PoS (SU_{wp}). To simplify the analysis, for SU_{wp} we just took the first guess from the systems on the **best** task (*bg*) that is used for scoring **best** mode recall (*Mode R*).

We then calculated the average intersection between these two sets:

$$TypeOverlap = \frac{\sum_{wp \in WP} |GU_{wp} \cap SU_{wp}|}{|WP|} \tag{8}$$

where *WP* is the full set of word and PoS combinations in the test data. This figure is given in the second column of Table 9. We also calculated the average number of substitutes types per word and PoS combination over the ten sentences for each system, again using the first response (*bg*) of the system. This calculation is shown as #subs in Table 9 and reflects the variation in responses i.e. did the system vary the best guess or was a given candidate often applied in many different contexts. The final column in Table 9 ($TypeOverlap_{uniq}$) is the average number of $|SU_{wp} \cap GU_{wp}|$ that were not found in the responses from ANY of the other systems. This demonstrates the original contribution of the system in finding good candidates that other systems failed to find.

Performance on these metrics is related to the resource used, rather than the correct differentiation of contexts, but it is the variation of the response over the sentences which will determine the score because we are not considering the token frequency of annotator responses, but the union of the substitute types. Both KU and USYD systems obtain a higher score on $Type\ Overlap_{uniq}$ reflecting the fact that they both use resources for obtaining substitutes (Roget and the Web 1T) that are not used by other systems in the **best** subtask. From the # subs we see that systems such as KU, USYD and IRST2 were quite varied in their responses, with USYD providing the most variation. These systems also all used the Web 1T corpus for matching context to substitute, albeit in different ways. The UNT system also used this corpus for n-gram modelling, but alongside many other resources. While UNT did not show such a varied response, it was clearly adept at matching context to candidate given its high ranking on the **best** and **oot** recall metrics. We shall turn to the disambiguation capabilities of systems in the next subsection.

4.3 Disambiguation analysis

In this section we analyse the results so as to demonstrate the disambiguation capabilities of the systems for sentences where they have found the **best** mode within SU_{wp}. We do this in order to focus on disambiguation, rather than whether

[14] We do not further complicate this analysis by considering the frequency of these responses.

Table 9 Overlap analysis of the system's responses with responses from the gold-standard on the **best** task for word and PoS combinations

System	TypeOverlap	#subs	$TypeOverlap_{uniq}$
KU	2.88	6.30	0.58
USYD	2.58	7.51	0.54
IRST2	2.57	5.50	0.29
MELB	1.91	3.77	0.27
HIT	1.87	4.29	0.18
IRST1	1.65	4.22	0.35
UNT	1.60	2.90	0.30
TOR	0.70	3.66	0.14

the system came up with the substitute as a candidate. We calculate precision using the *bg* compared to the mode from the annotators on:

All Items where all systems found the mode within their SU_{wp} (NB there were only 17 such items)

Sys Items where the given system found the mode within its SU_{wp}

That is, precision is calculated as:

$$All\ precision = \frac{\sum_{bg_i \in All} 1\ if\ bg_i = m_i}{|All|} \tag{9}$$

and

$$Sys\ precision = \frac{\sum_{bg_i \in Sys} 1\ if\ bg_i = m_i}{|Sys|} \tag{10}$$

As in the previous subsection, we restrict ourselves to items where the target word was not part of a multiword and we use the first single word guess as best guess.

The results are displayed in Table 10. As predicted, given the high scores for the task yet low variation per candidate, UNT shows optimal performance at disambiguation for items where it had found the candidate within SU_{wp}. Typically systems with more variation will have a harder disambiguation task on items where the mode is within their SU_{wp}. The other systems that use a search engine query for disambiguation (HIT and MELB) also do well on **Sys** precision. The systems TOR and IRST1 that restricted training to the BNC performed lower than those that used either the Web 1T or a search engine for disambiguation. It would be interesting to know if

Table 10 Precision against mode on items where (i) All systems (ii) this Sys(tem) found the mode within all SU_{wp}

System	HIT	UNT	KU	MELB	USYD	IRST2	IRST1	TOR
All	58.82	52.94	52.94	47.06	47.06	41.18	35.29	23.53
Sys	52.53	59.67	42.31	53.71	37.77	44.57	43.82	37.91

they would have outperformed an n-gram approach, such as KU, if the n-grams were obtained from the BNC. The USYD approach also does not do as well on **Sys**. This may be because it has to choose between a great many candidates (see the #subs column in Table 9) or it may be because the *pMI* filtering it performs on the Web 1T does not provide such good disambiguation as the n-gram approach.

4.4 Post hoc analysis

The task of choosing a lexical substitute for a given word is not clear cut and there is inherent variation in the task. It is likely that there will be synonyms that humans do not think of. We therefore conducted a post hoc analysis after obtaining the results from the systems to check 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 two single word substitutes, no NAME responses, and where the target word was not one of those identified as a multiword by two or more annotators. We then mixed the substitutes from the human annotators with those provided by the systems for the **best** task. Three fresh annotators were given the test sentences and asked to categorise the randomly ordered substitutes for each sentence as good, reasonable or bad. These post hoc annotators were all native English speakers from the UK. When collating the post hoc responses, we took the majority verdict for each substitute. 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 majority vote of the post hoc annotators are shown in Table 11. 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.

Table 12 gives the percentage of substitute types under each category where the substitute was found by: (i) ONLY original annotators (ii) ONLY systems (iii) both. This measures how much the systems find substitutes that the humans did not think of, and vice versa. From these figures we can verify that systems are capable of

Table 11 Post hoc results

	Total	Good	Reasonable	Bad
sys	1158	9.07	19.08	71.85
origA	356	37.36	41.01	21.63

Table 12 Post hoc verdicts on substitutes given only by system, original annotators or those given by both

	Total	Good	Reasonable	Bad
sys only	985	3.05	15.03	81.93
origA only	183	31.69	39.89	28.42
both	173	43.35	42.20	14.45

generating many more substitutes than humans but these are not typically of such high quality. Only a very small percentage of responses from the systems ONLY are liked by the post hoc annotators. This is a difficult task without cut and dried answers and the original annotators do not always provide responses that are liked by the post hoc annotators, however the proportion of disliked substitutes is far less considering the substitutes only from humans (28%) compared to those only from systems (82%). If we look at these examples where only humans provided the substitute and the majority of post hoc annotators categorised the human annotation as ‘bad’ we find many that look reasonable to us, for example:

Appointed by the CDFA, public members are chosen for their usefulness in helping the commodity <u>board</u> carry out its purpose and to represent the public interest.

The annotation judged as “bad” was *management* which seemed reasonable to us.

There were 52 such substitutions[15] from the original human annotators (not systems) deemed “bad” by the post hoc annotators. Of these cases, 50 were substitutes provided by only one annotator of the five, and the remaining two cases involved substitutes provided by only two of the original annotators. Furthermore, in 38 of these 52 occurrences one of the three post hoc annotators was of a different opinion: 31 times the outlier annotator selected “reasonable” and 7 times they selected “good” (as in the *board* example above). While there are still 14 substitutes provided only by humans which all post hoc annotators disliked, it is much easier to judge the annotations of others, than to find substitutes from scratch. In both the cases where the substitute was selected by two of the original annotators, at least one of the post hoc annotators gave a “reasonable” verdict.

5 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 and different choices might be equally valid even though they were not selected by our human annotators,[16] though as we show from the post hoc evaluation, only 18% (3% good and 15% reasonable) of substitutes from the systems that were considered good or reasonable were not elicited by our original five annotators. The inherent variation 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 usage of words to convey 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.

[15] Please note that whilst the post hoc evaluation looked at 100 sentences, there were 1342 substitutes in total for these 100 sentences examined by post hoc annotators.

[16] This could be rectified to some extent by recruiting more annotators, possibly using volunteers in a web-based game (Mihalcea and Chklovski 2003).

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 first author from the UK EPSRC project EP/C537262/1 "Ranking Word Senses for Disambiguation: Models and Applications" and to the second author from INTEROP NoE (508011, 6th EU FP). We thank the annotators for their hard work, the anonymous reviewers for their useful feedback, Serge Sharoff for the use of his Internet corpus, Julie Weeds for the distributional similarity software and Suzanne Stevenson for suggesting the **oot** task.

References

Barnard, J. (Ed.) (1986). *Macquarie Thesaurus*. Sydney: Macquarie Library.

Brants, T., & Franz, A. (2006). Web 1T 5-gram corpus version 1.1. Technical Report.

Briscoe, E., & Carroll, J. (2002). Robust accurate statistical annotation of general text. In *Proceedings of the third international conference on Language Resources and Evaluation (LREC)* (pp. 1499–1504). Las Palmas, Canary Islands, Spain.

Carpuat, M., & Wu, D. (2007). Improving statistical machine translation using word sense disambiguation. In *Proceedings of the joint conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)* (pp. 61–72). Prague, Czech Republic.

Chan, Y. S., Ng, H. T., & Chiang, D. (2007). Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics* (pp. 33–40). Prague, Czech Republic.

Dagan, I., Glickman, O., & Magnini, B. (2005). The PASCAL recognising textual entailment challenge. In *Proceedings of the PASCAL first challenge workshop* (pp. 1–8). Southampton, UK.

Dahl, G., Frassica, A.-M., & Wicentowski, R. (2007). SW-AG: Local context matching for English lexical substitution. In *Proceedings of the 4th workshop on Semantic Evaluations (SemEval-2007)* (pp. 304–307). Prague, Czech Republic.

Fellbaum, C. (Ed.) (1998). *WordNet, an electronic lexical database*. Cambridge, MA: The MIT Press.

Giuliano, C., Gliozzo, A., & Strapparava, C. (2007). FBK-irst: Lexical substitution task exploiting domain and syntagmatic coherence. In *Proceedings of the 4th workshop on Semantic Evaluations (SemEval-2007)* (pp. 145–148). Prague, Czech Republic.

Graff, D. (2003). *English Gigaword*. Philadelphia: Linguistic Data Consortium.

Hanks, P. (2000). Do word meanings exist?. *Computers and the Humanities. Senseval Special Issue, 34*(1–2), 205–215.

Hassan, S., Csomai, A., Banea, C., Sinha, R., & Mihalcea, R. (2007). UNT: SubFinder: Combining knowledge sources for automatic lexical substitution. In *Proceedings of the 4th workshop on Semantic Evaluations (SemEval-2007)* (pp. 410–413). Prague, Czech Republic.

Hawker, T. (2007). USYD: WSD and lexical substitution using the Web1T corpus. In *Proceedings of the 4th workshop on Semantic Evaluations (SemEval-2007)* (pp. 446–453). Prague, Czech Republic.

Ide, N., & Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics, 24*(1), 1–40.

Ide, N., & Wilks, Y. (2006). Making sense about sense. In E. Agirre & P. Edmonds (Eds.), *Word sense disambiguation, algorithms and applications* (pp. 47–73). Springer.

Kilgarriff, A. (2004). How dominant is the commonest sense of a word? In *Proceedings of text, speech, dialogue*. Brno, Czech Republic.

Kilgarriff, A. (2006). Word senses. In E. Agirre & P. Edmonds (Eds.), *Word sense disambiguation, algorithms and applications* (pp. 29–46). Springer.

Lee, L. (1999). Measures of distributional similarity. In *Proceedings of the 37th annual meeting of the association for computational linguistics* (pp. 25–32).

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

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th international conference on machine learning*. Madison, WI.

Lindberg, C. (Ed.) (2004). *The Oxford American Writer's Thesaurus*. Oxford: Oxford University Press.

Martinez, D., Kim, S. N., & Baldwin, T. (2007). MELB-MKB: Lexical substitution system based on relatives in context. In *Proceedings of the 4th workshop on Semantic Evaluations (SemEval-2007)* (pp. 237–240). Prague, Czech Republic.

McCarthy, D. (2002). Lexical substitution as a task for WSD evaluation. In *Proceedings of the ACL workshop on word sense disambiguation: Recent successes and future directions* (pp. 109–115). Philadelphia, USA.

McCarthy, D. (2008). Lexical substitution as a framework for multiword evaluation. In *Proceedings of the sixth international conference on Language Resources and Evaluation (LREC 2008).* Marrakech, Morocco.

McCarthy, D., & Navigli, R. (2007). SemEval-2007 Task 10: English lexical substitution task. In *Proceedings of the 4th workshop on Semantic Evaluations (SemEval-2007)* (pp. 48–53). Prague, Czech Republic.

Mihalcea, R., & Chklovski, T. (2003). Open mind word expert: Creating large annotated data collections with Web Users' help. In *Proceedings of the EACL 2003 workshop on linguistically annotated corpora* (pp. 53–60). Budapest.

Mihalcea, R., & Csomai, A. (2005). SenseLearner: Word sense disambiguation for all words in unrestricted text. In *Proceedings of the 43rd annual meeting of the association for computational linguistics*. University of Michigan, USA.

Miller, G. A., Leacock, C., Tengi, R., & Bunker, R. T. (1993). A semantic concordance. In *Proceedings of the ARPA workshop on human language technology* (pp. 303–308).

Mohammad, S., Hirst, G., & Resnik, P. (2007). Tor, TorMd: Distributional profiles of concepts for unsupervised word sense disambiguation. In *Proceedings of the 4th workshop on Semantic Evaluations (SemEval-2007)* (pp. 326–333). Prague, Czech Republic.

Navigli, R. (2006). Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the 44th annual meeting of the association for Computational Linguistics joint with the 21st International Conference on Computational Linguistics (COLING-ACL 2006)* (pp. 105–112). Sydney, Australia.

Navigli, R., Litkowski, K. C., & Hargraves, O. (2007). SemEval-2007 Task 7: Coarse-grained English all-words task. In *Proceedings of the 4th workshop on Semantic Evaluations (SemEval-2007)* (pp. 30–35). Prague, Czech Republic.

Palmer, M. (2000). Consistent criteria for sense distinctions. *Computers and the Humanities. Senseval Special Issue, 34*(1–2), 217–222.

Palmer M., Dang, H. T., & Fellbaum, C. (2007). Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering, 13*(02), 137–163.

Pantel, P., & Lin, D. (2002). Discovering word senses from text. In *Proceedings of ACM SIGKDD conference on knowledge discovery and data mining* (pp. 613–619). Edmonton, Canada.

Resnik, P., & Yarowsky, D. (2000). Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering, 5*(3), 113–133.

Roget, P. M. (1911). *Roget's International Thesaurus* (1st ed.). New York, USA: Cromwell.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics, 24*(1), 97–123.

Schütze, H., & Pederson, J. O. (1995). Information retrieval based on word senses. In *Proceedings of the fourth annual symposium on document analysis and information retrieval* (pp. 161–175). Las Vegas, NV.

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

Stokoe, C. (2005). Differentiating homonymy and polysemy in information retrieval. In *Proceedings of the joint conference on human language technology and empirical methods in natural language processing* (pp. 403–410). Vancouver, BC, Canada.

Stokoe, C., Oakes, M. P., & Tait, J. (2003). Word sense disambiguation in information retrieval revisited. In *Proceedings of SIGIR* (pp. 159–166).

Thesaurus.com. (2007). *Roget's New Millennium*$^{\mathrm{TM}}$ *Thesaurus*(ist ed., v 1.3.1). Lexico Publishing Group, LLC. http://thesaurus.reference.co.

Yuret, D. (2007). KU: Word sense disambiguation by substitution. In *Proceedings of the 4th workshop on Semantic Evaluations (SemEval-2007)* (pp. 207–214). Prague, Czech Republic.

Zhao, S., Zhao, L., Zhang, Y., Liu, T., & Li, S. (2007). HIT: Web based scoring method for English lexical substitution. In *Proceedings of the 4th workshop on Semantic Evaluations (SemEval-2007)* (pp. 173–176). Prague, Czech Republic.