

# Word Sense Disambiguation with an Integrated Lexical Resource

Oi Yee Kwong

Language Information Sciences Research Centre  
City University of Hong Kong  
Tat Chee Avenue, Kowloon, Hong Kong  
rlolivia@cityu.edu.hk

## Abstract

Word sense disambiguation (WSD) notably requires a lot of different types of lexical information to perform well. In this paper, we report a WSD study using a lexical resource which is an integration of WordNet 1.6 and Roget's Thesaurus. Various types of information were employed from this resource for WSD. Experimental results indicate that different types of information may be most useful for disambiguating words in different text types. More importantly, the results demonstrate the practicality of the integrated resource, and thus how WSD can benefit from using enriched lexical information by integrating existing lexical resources.

## 1 Introduction

Word sense disambiguation (WSD) notably requires a wide range of lexical information to perform well. To form a comprehensive lexical resource, one may start from scratch and encode as much information as desired into a lexical database manually (e.g. Small and Rieger, 1982). Alternatively, one may acquire lexical semantic information semi-automatically from existing resources such as machine-readable dictionaries (e.g. Amsler, 1981; Chodorow et al., 1985). However, as suggested by Kwong (1998; 2001), the most cost-effective way by far appears to be combining various existing resources, especially different types of them which are exemplars for representing different types of information, into an integrated repository. Although such resource integration has been suggested (e.g. Yarowsky, 1992) or implemented in some way (e.g. Knight and Luk, 1994; McHale and Crowter, 1994), surprisingly few studies have actually made use of multiple existing lexical resources simultaneously in WSD.

McRoy (1992) is among the first to seriously investigate the use of multiple types of information in WSD. She used several types of knowledge including syntactic tag, morphology, collocation, word association, and selectional preference for WSD. The information was distributed across a core lexicon (for general, coarse sense distinctions), a dynamic lex-

icon (for context-specific, finer sense distinctions), a concept hierarchy, a set of collocation patterns, and a set of cluster definitions. Subsequent studies (e.g. Bruce and Wiebe, 1994; Ng and Lee, 1996; Harley and Glennon, 1997; Leacock et al., 1998; Wilks and Stevenson, 1998) used different combinations of more or less similar types of knowledge as those used by McRoy, while the information was derived from different resources. Although varied results were reported, all these tests provided some support for the use of multiple knowledge for WSD.<sup>1</sup> In fact, a hybrid system might be necessary because words differ in many respects including their number of senses, frequency of these senses, availability of training samples and so on (Hawkins and Nettleton, 2000).

Nevertheless, most studies which use multiple knowledge sources for WSD are only maximising the exploitation of a single (type of) resource. McRoy (1992) may be an exception, but her resources were tailor-made and the linkage between them was manually imposed. So rather than using a "standard" lexicon for WSD, in this study we make use of an *Integrated Lexical Resource* (ILR), which was the result of automatically linking WordNet 1.6 and Roget's Thesaurus. We obtain a range of semantic relations of different specificity from the ILR and apply them in WSD, to demonstrate the practicality of the ILR in providing enriched information for WSD.

In Section 2, we will briefly introduce the ILR. Then in Section 3, we will describe a spectrum of semantic relations, i.e. semantic relations of various degrees of specificity, and how such relations are obtained from the ILR. We will test the individual and combined effect of these semantic relations on WSD in Sections 4 and 5 respectively. An example illustrating the results is given in Section 6, followed by a conclusion in Section 7.

---

<sup>1</sup>Accuracy figures from 54% (Ng and Lee, 1996) to over 80% (e.g. Bruce and Wiebe, 1994; Leacock et al., 1998) were seen. We believe the variation is largely due to the difference in sense distinction and test materials.

## 2 The Integrated Lexical Resource

By *integration*, we mean that the various resources are linked in some way but the different types of information are preserved in their original structures in individual resources.

WordNet (Miller et al., 1990) groups words into sets of synonyms (“synsets”), with an optional textual gloss. These synsets form the nodes of a taxonomic hierarchy. Roget’s Thesaurus, on the other hand, groups words into approximately 1,000 semantic heads. Under each head, words are first assorted by part of speech and then grouped into paragraphs according to broad conceptual categories. Heads are in turn grouped into sections and then into classes, resulting in a hierarchical organisation. Thus WordNet and Roget’s Thesaurus are exemplars for representing different types of information, namely narrow and broad semantic relations respectively.

Kwong (2001) used a structurally-based sense-mapping algorithm to perform a full-scale linkage of the noun senses in common to WordNet 1.6 (WN16) and Roget’s Thesaurus (ROGET).<sup>2</sup> Although the two resources differ considerably in collection size, about 90% of the nouns shared by both resources could be linked by the algorithm in some way.<sup>3</sup> The full-scale linkage resulted in more than 18,000 nouns and 30,000 senses being mapped, forming the Integrated Lexical Resource used in the current study. Although it had been difficult and impractical to check the mappings in the ILR exhaustively given the huge amount of data, extensive sampling of the results showed that over 70% of the mappings are expected to be accurate (Kwong, 1998; Kwong, 2001).

It happens that both component resources in the ILR have a hierarchical classification of senses. Although they follow quite different classificatory criteria, they can now be used compatibly within the ILR. Apart from the information originally available from the individual component resources, extra information could be obtained with the resultant resource, as we will see in the next section.

## 3 A Spectrum of Semantic Relations

In this section, we will describe the semantic relations used in this study and how they are obtained computationally from the ILR. We organise the semantic relations in terms of their specificity, from the narrowest type to the broadest type. This ranking is essentially intuitive, taking also the number of component lexical resources involved into account.

<sup>2</sup>Both the 1911 Crowell Roget’s Thesaurus and the 1987 Penguin Roget’s Thesaurus were used.

<sup>3</sup>WN16 has 94,474 nouns with 116,364 senses, while ROGET has 35,718 nouns with 75,013 senses.

### 3.1 Narrow Semantic Relation (*wn*)

This is the relation captured in WordNet 1.6 and we measured the similarity between two senses  $x$  and  $y$  by taxonomic distance with the simple metric from Rada et al. (1989):

$$S(x, y) = \frac{1}{1 + \text{Dist}(x, y)}$$

where  $\text{Dist}(x, y)$  is the minimum number of edges separating the nodes representing  $x$  and  $y$ .

### 3.2 Broad Semantic Relation 1 (*rog*)

The similarity of two WordNet senses  $x$  and  $y$  with respect to the Roget classification (now a component of the ILR) is measured by the following empirically determined function:

$$S(x, y) = \begin{cases} |M(x) \cap M(y)| & \text{if } M(x) \cap M(y) \neq \emptyset \\ & \text{and } (M(x) \cap M(y)) \cap \text{TopC} \neq \emptyset \\ 0.25 & \text{if } M(x) \cap M(y) = \emptyset \\ & \text{and } h_1 \in M(x), h_2 \in M(y), \\ & h_3 \in \text{TopC}, S(h_1) = S(h_2) = S(h_3) \\ 0 & \text{otherwise} \end{cases}$$

where

$M(s)$  is the set of Roget heads that sense  $s$  is mapped to,

$S(h)$  is the Roget section (one level above heads) subsuming head  $h$ , and

$\text{TopC}$  is the set of Roget heads covering most words in the text. Heads at the top two positions with eight or more words in the text are included (determined empirically), and ties are preserved.

### 3.3 Broad Semantic Relation 2 (*def*)

This is the word overlap between pairs of sense definitions in the tradition of Lesk (1986). Sense definitions are based on the WordNet glosses, with stop words and duplicated words removed.<sup>4</sup> The similarity between senses is measured by the number ( $\geq 2$ ) of words common to their definitions, normalised with the Dice function:

$$S(x, y) = \begin{cases} \frac{2|X \cap Y|}{|X| + |Y|} & \text{if } |X \cap Y| > 1 \\ 0 & \text{otherwise} \end{cases}$$

where

$X$  is the definition of sense  $x$ , and

$Y$  is the definition of sense  $y$ .

### 3.4 Broad Semantic Relation 3 (*com*)

This relation directly measures the relevance of a sense to the text as a whole. Each sense (in the text) is assigned a score based on the length of overlap between its WordNet definition and the set of words covered by the Roget heads in  $\text{TopC}$  (as defined in Section 3.2).

<sup>4</sup>Lacking reliable mechanisms, no stemming was done. However, perfect stemming should be beneficial for definition matching.

## 4 Experiment 1

### 4.1 Materials and Method

Test materials were taken from the SEMCOR data, which consist of texts from the Brown Corpus tagged with WordNet senses. The texts are about 2,000 words in length each. There are 186 texts, divided into 15 categories, amounting to more than 60,000 test instances for more than 5,000 distinct nouns, with number of senses ranging from 2 to 30, average at 3.41. Each polysemous noun occurrence (lemmatized) was treated as a target word.

The test files were disambiguated with each of the different semantic relations described in Section 3. For comparison, we also tagged the nouns with their first sense from WN16, which is expected to be the most frequent sense (*freq*).

Apart from *com* which was run as a one-step heuristic, others (*wn*, *rog* and *def*) were run with a recursive filtering algorithm driven by monosemous nouns from the whole text.<sup>5</sup> We start by comparing all monosemous nouns with all other polysemous nouns within a text (with the various measures of  $S(x, y)$  in Section 3). All senses with scores over a certain threshold are filtered in subsequent rounds, to be compared with the rest of the polysemous nouns. The algorithm converges when the threshold reaches a minimum and no more senses are filtered (Kwong, 2000).

### 4.2 Performance Measures

The principal measures of performance are *precision* and *recall*, which we define for this study as follows:

$$\textit{precision} = \frac{\textit{Corr\_DTS}}{\textit{DTS}}$$
$$\textit{recall} = \frac{\textit{Corr\_DTS}}{\textit{Total}}$$

where *Corr\_DTS* is the number of correctly disambiguated test samples, *DTS* is the number of disambiguated test samples, and *Total* is the number of test samples.

### 4.3 Results and Discussion

Table 1 shows the average precision and recall from different semantic relations. The best outcomes (excluding *freq*) in each row are shaded.

On average, *wn* and *rog* both achieve over 40% precision. However, there is a nearly two-fold difference in recall. The inferiority of *rog* can be attributed to two factors. First, ROGET has a much smaller collection size than WN16, and the number

of words shared by them is even smaller. Second, the results depend on not only the actual effectiveness of the semantic relation but also the accuracy of the sense linkage between WN16 and ROGET, which was about 70% as mentioned. However, *wn* and *rog* have similar precision, suggesting that where the information is applicable, both types of information are similarly effective for disambiguation.

Meanwhile, the other broad semantic relations involving the definitions of the senses, i.e. *def* and *com*, appear to result in lower precision than *wn* and *rog* but comparable recall. This is probably because definitions contain “noisy” words, although an effort has been made to keep them to a minimum (e.g. by removing stop words).

As for *freq*, it gives almost 70% precision and recall on average. Compared to other studies which used most frequent sense / first sense as baseline, what we get appears extraordinarily high. For instance, Wilks and Stevenson (1998) got a baseline of only 49.8% using first sense in the Longman Dictionary of Contemporary English. We believe that the anomaly of *freq* is mostly due to the circularity between SEMCOR and WordNet: the former is tagged with senses from the latter while the sense ordering in the latter at least partly depends on what is observed from the former.

The results in Table 1 also suggest that narrow semantic relations may not be a very suitable type of information for disambiguating words in fictional texts. In terms of recall, for instance, while *wn* does not work well for the fictional categories including K, L, M, N, and P, these categories are all best disambiguated by *com*, the broadest semantic relation according to our ranking in Section 3. It was observed that different text types have considerably different noun compositions. For instance, comparing files br-j03 (scientific writing) and br-k22 (fiction), nouns are far more repetitive in the former than in the latter. Moreover, nouns found in br-j03 are more likely to be highly “topical” in Leacock et al.’s (1998) sense, but those found in br-k22 are relatively more domain-neutral and multi-purpose. This difference is also reflected by the monosemous nouns in the texts, such as “velocity” and “manometer” in br-j03, and “stranger” and “desk” in br-k22. While further analysis would be required to justify the relation between semantic relations and text types, the observed contrast in the present experiment nevertheless suggests that different types of semantic relation are needed as they may complement one another at different situations (e.g. text types), leading us to the next experiment on the use of multiple types of information for WSD.

<sup>5</sup>To compensate for the possible inadequacy of the sense linkage, we also assigned credits to the candidate senses which were under the ROGET heads in *TopC*, in case the mapping has masked some likely senses and thus prevented them from being filtered.

Text Category	Precision (%)					Recall (%)				
	<i>wn</i>	<i>rog</i>	<i>def</i>	<i>com</i>	<i>freq</i>	<i>wn</i>	<i>rog</i>	<i>def</i>	<i>com</i>	<i>freq</i>
A (Press:Reportage)	56.90	38.14	42.35	39.30	69.33	37.86	12.47	29.55	29.82	69.33
B (Press:Editorial)	40.63	38.31	30.23	27.65	64.99	29.46	12.14	19.38	19.02	64.99
C (Press:Reviews)	40.27	46.96	28.72	33.07	63.37	27.31	21.24	19.96	24.48	63.37
D (Religion)	40.35	47.59	31.65	33.31	63.27	27.88	14.46	19.59	24.41	63.27
E (Skills and Hobbies)	46.75	45.08	38.83	39.93	67.74	35.49	18.06	27.07	29.55	67.76
F (Popular Lore)	46.60	46.44	38.31	40.29	67.36	34.77	17.49	26.94	30.29	67.36
G (Belles Lettres, Biography, etc.)	39.80	42.12	32.20	34.03	63.69	30.17	16.24	21.92	25.16	63.69
H (Miscellaneous)	45.61	44.70	38.62	43.26	71.08	34.30	16.75	25.62	32.04	71.08
J (Learned and Scientific Writing)	44.04	42.34	38.89	38.85	69.97	32.94	16.59	23.11	28.96	69.97
K (Fiction:General)	36.32	38.55	31.25	33.68	71.34	21.26	12.40	17.61	23.32	71.35
L (Fiction: Mystery and Detective)	34.60	41.84	25.58	35.44	67.33	19.35	13.05	12.68	26.46	67.35
M (Fiction:Science)	37.83	54.87	34.93	41.69	69.18	23.43	15.34	21.39	28.41	69.31
N (Fiction:Adventure and Western)	37.88	37.20	30.77	34.04	71.65	20.54	14.79	15.40	25.79	71.65
P (Fiction:Romance and Love Story)	42.88	44.96	32.46	30.70	69.78	22.46	13.59	18.81	23.30	69.78
R (Humour)	35.27	37.79	29.09	29.95	69.53	23.37	15.29	16.94	22.38	69.57
<b>Average</b>	42.02	42.28	34.90	36.81	68.79	29.05	15.50	21.68	27.15	68.80

Table 1: Results of Experiment 1

## 5 Experiment 2

### 5.1 Materials and Method

The test materials (i.e. 186 SEMCOR files) and method (i.e. application of different types of information) were the same as in Experiment 1. As mentioned in Section 1, most WSD studies which use multiple types of information usually constrain the information within a single resource. Few have actually used multiple lexical resources simultaneously. We will thus cover both possibilities in this experiment: for multiple types of information within a single resource, we tested *wn+def*; for multiple types of information from multiple resources, we tested *wn+rog*, *wn+rog+com*, and *wn+rog+com+def*. Moreover, for all these combinations, two variations – with and without *freq* (*+freq* and *-freq* respectively) – were tested.

The actual scores from individual information types were first converted to *relative scores* to normalise the different scoring scales as follows:

$$S_r(w_i) = \frac{S_a(w_i)}{\max_j(S_a(w_j))}$$

where  $S_r(w_i)$  is the relative score for the  $i^{th}$  sense of word  $w$ , and  $S_a(w_i)$  is the *actual score* for the  $i^{th}$  sense of word  $w$ . The relative scores from different information types for each sense were then added up for the various information combinations correspondingly. The measures of performance were precision and recall as in the last experiment.

### 5.2 Results and Discussion

The overall average precision (P) and recall (R) for the various combinations of information types are shown in Table 2.

Information Combination		P (%)	R (%)
<i>wn+def</i>	<i>-freq</i>	43.24	35.50
	<i>+freq</i>	67.97	66.35
<i>wn+rog</i>	<i>-freq</i>	44.40	33.97
	<i>+freq</i>	70.82	68.91
<i>wn+rog+com</i>	<i>-freq</i>	42.59	37.65
	<i>+freq</i>	65.14	62.93
<i>wn+rog+def+com</i>	<i>-freq</i>	42.13	38.63
	<i>+freq</i>	61.22	59.36

Table 2: Overall Results of Experiment 2

Comparing Tables 1 and 2, we see that all combinations of semantic relations (at *-freq*) outperform their individual component relation (e.g. *wn+def* performs better than *wn* or *def* alone). However, the effect of information combination is not necessarily additive. In other words, more types of information do not always give proportionally better results. For example, *wn+rog+def+com* only gives a very minor increase in recall compared to *wn+rog+com*. This situation is especially obvious when *freq* is included. For instance, given the extraordinary performance of *freq* alone in Experiment 1, we found in Experiment 2 that only *wn+rog+freq* outperforms *freq* on its own. The recall in the former is marginally better and the precision is 2 to 3% higher. An example showing how *wn* and *rog* combined and the overlapping between them will be discussed in Section 6.

There are at least two possible causes for the above results regarding the combination of information types. On the one hand, two or more different types of information may in fact be good for disambiguating a similar set of words, so they do not offer much additional advantage to WSD when combined. On the other hand, the effect of different

types of information may offset one another when indiscriminately combined. That means one type of information may be good for disambiguating particular words while another type of information may be adverse for those words. Therefore, for the combination of information to be effective, conflicting evidence must be resolved. Handling conflicting evidence thus requires the preferences of the target words (and their senses) to be known, especially if WSD is, as Resnik and Yarowsky (1999) suggested, highly lexically sensitive.

The results in this experiment suggest that using multiple types of information for WSD is often beneficial to WSD. In particular, narrow and broad semantic relations are most useful in different text categories and may therefore complement each other. Most importantly, we have demonstrated that the two types of semantic relation can be effectively obtained and used together by integrating two existing lexical resources, despite the limited success of that integration at present. In the next section, we will look more closely at an example to further illustrate the results.

## 6 An Example

Table 3 shows part of the results for test file br-j03, a 1944-word (with 461 noun occurrences, of which 100 are monosemous) scientific writing describing some physics experiment of fluid using a band viscometer. The benefit of using multiple types of information from the ILR is evident. Combining *wn* and *rog* gives precision and recall which are superior to either *wn* or *rog* alone. Moreover, adding *freq* to *wn+rog* outperforms *freq* itself.

Information		P (%)	R (%)
Narrow Semantic Relation ( <i>wn</i> )		47.02	39.34
Broad Semantic Relation ( <i>rog</i> )		57.98	38.23
Frequency ( <i>freq</i> )		65.65	65.65
Combined ( <i>wn+rog</i> )	- <i>freq</i>	58.58	54.85
	+ <i>freq</i>	74.08	72.85

Table 3: WSD Results for br-j03

To see the contribution of individual types of information, consider the actual words being correctly disambiguated, as shown in Figure 1. Senses marked with \* cease to be correct with *wn+rog*, whereas those marked with  $\star$  were only correct with *wn+rog*. The number after the slash (/) indicates the expected sense with respect to WN16.

We can see that there are some words which can be disambiguated with either resource (e.g. axis, energy), but more are exclusively disambiguated by information from one resource (e.g. tape, strain, and entropy with *rog*, and field, tension, and poise with *wn*). Using narrow and broad semantic relations together can sometimes overcome the “tennis

problem” mentioned in Fellbaum (1998). For example, energy/1, force/1 and suction/2 are all close to heat/1 and field/5 in WN16 for being “physical phenomena”. Although shear/1 and strain/1 are also obviously physics-related, they are classified as deformation, i.e. event, in WN16. Nevertheless, ROGET groups the latter two senses together with force, energy, and suction from another perspective, i.e. power (head 160), thus enabling them to be disambiguated.

Correct disambiguation can also result from evidence accumulation. For example, “stress”, which is wrongly tagged with either *wn* or *rog* alone, is correctly disambiguated with *wn+rog*. The fact is that stress/5 (the expected sense) lost to different senses with *wn* and *rog* alone, but with *wn+rog*, the correct sense ended up with most accumulated evidence.

## 7 Conclusion

We have thus studied the individual and combined effect of semantic relations of varied specificity on WSD, with all the relations obtained from the Integrated Lexical Resource. It was found that different words seemed to be better disambiguated with different types of semantic relation, and this interaction was closely related to text types. We have shown that different types of information might complement one another and demonstrated the practicality of the ILR. More importantly, we have shown how WSD can benefit from integrating existing lexical resources, which enables different types of lexical information to be applied simultaneously, flexibly, and compatibly.

## Acknowledgements

This work was done in the Computer Laboratory, University of Cambridge. The author would like to thank Prof. Karen Sparck Jones for her advice and comments. The work was financially supported by the Committee of Vice-Chancellors and Principals of the Universities of the United Kingdom, and the Croucher Foundation.

## References

- R. Amsler. 1981. A taxonomy for English nouns and verbs. In *Proceedings of the 19th Annual Meeting of the Association for Computational Linguistics (ACL '81)*, pages 133–138, Stanford.
- R. Bruce and J. Wiebe. 1994. Word-sense disambiguation using decomposable models. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL '94)*, pages 139–145, Las Cruces, New Mexico.
- M.S. Chodorow, R.J. Byrd, and G.E. Heidorn. 1985. Extracting semantic hierarchies from a large online dictionary. In *Proceedings of the 23rd Annual*

<u>wn</u>	<u>wn/rog</u>	<u>rog</u>
angle/1, area/6, assumption/2, basis/2, change/2, concentration/2, conclusion/2, diameter/1, difference/2, *equation/1, field/5, fluid/2, heat/1, height/1, hole/4, *inclination/3, instrument/1, manner/1, material/1, measurement/1, mechanics/1, part/1, plane/2, poise/1, *property/2, relation/1, relationship/1, shortness/1, statement/1, temperature/1, tension/5, theory/2, type/1, unit/1, upper.limit/1, use/1, *value/1, way/1, *work/1	axis/1, energy/1, force/2, head/6, pressure/1, rate/2, reason/2, suction/1, system/7	*amount/2, band/6, breakup/2, calculation/1, chance/2, chemist/1, design/1, direction/2, effect/1, entropy/1, exception/1, *experiment/2, literature/1, paper/2, run/2, shear/1, *solution/1, strain/1, tape/1, *volume/1, work/7
	*stress/5	

Figure 1: Words Correctly Disambiguated in br-j03

- Meeting of the Association for Computational Linguistics (ACL '85)*, pages 299–304, Chicago.
- C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA.
- A. Harley and D. Glennon. 1997. Sense tagging in action: Combining different tests with additive weightings. In *Proceedings of SIGLEX '97 Workshop: Tagging Text with Lexical Semantics: Why, What, and How?*, pages 74–78, Washington, D.C.
- P. Hawkins and D. Nettleton. 2000. Large scale WSD using learning applied to SENSEVAL. *Computers and the Humanities: Special Issue on SENSEVAL*, 34(1-2):135–140.
- K. Knight and S.K. Luk. 1994. Building a large-scale knowledge base for machine translation. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, pages 773–778, Seattle, Washington.
- O.Y. Kwong. 1998. Aligning WordNet with additional lexical resources. In *Proceedings of the Workshop on the Usage of WordNet in Natural Language Processing Systems*, pages 73–79, Montréal, Canada.
- O.Y. Kwong. 2000. Word sense selection in texts: An integrated model. Technical Report No.504, Computer Laboratory, University of Cambridge, U.K.
- O.Y. Kwong. 2001. Forming an integrated lexical resource for word sense disambiguation. In *Proceedings of the 15th Pacific Asia Conference on Language, Information and Computation (PACLIC 15)*, pages 109–119, Hong Kong.
- C. Leacock, M. Chodorow, and M.A. Miller. 1998. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24(1):147–165.
- M.E. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 1986 SIGDOC Conference*, pages 24–26, Toronto, Canada.
- M.L. McHale and J.J. Crowter. 1994. Constructing a lexicon from a machine readable dictionary. Technical Report RL-TR-94-178, Rome Laboratory, Griffiss Air Force Base, New York.
- S.W. McRoy. 1992. Using multiple knowledge sources for word sense disambiguation. *Computational Linguistics*, 18(1):1–30.
- G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K.J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.
- H.T. Ng and H.B. Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL '96)*, pages 40–47, Santa Cruz, CA.
- R. Rada, H. Mili, E. Bicknell, and M. Blettner. 1989. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30.
- P. Resnik and D. Yarowsky. 1999. Distinguishing systems and distinguishing senses: new evaluation methods for Word Sense Disambiguation. *Natural Language Engineering*, 5(2):113–133.
- S. Small and C. Rieger. 1982. Parsing and comprehending with word experts (A theory and its realization). In W.G. Lehnert and M.H. Ringle, editors, *Strategies for Natural Language Processing*. Lawrence Erlbaum Associates, New Jersey.
- Y. Wilks and M. Stevenson. 1998. Word sense disambiguation using optimised combinations of knowledge sources. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL '98)*, pages 1398–1402, Montréal, Canada.