Gloss-Based Semantic Similarity Metrics for Predominant Sense Acquisition

Ryu Iida
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0192, Japan
ryu-i@is.naist.jp

Diana McCarthy and **Rob Koeling**
University of Sussex
Falmer, East Sussex
BN1 9QH, UK
{dianam,robk}@sussex.ac.uk

Abstract

In recent years there have been various approaches aimed at automatic acquisition of predominant senses of words. This information can be exploited as a powerful back-off strategy for word sense disambiguation given the zipfian distribution of word senses. Approaches which do not require manually sense-tagged data have been proposed for English exploiting lexical resources available, notably WordNet. In these approaches distributional similarity is coupled with a semantic similarity measure which ties the distributionally related words to the sense inventory. The semantic similarity measures that have been used have all taken advantage of the hierarchical information in WordNet. We investigate the applicability to Japanese and demonstrate the feasibility of a measure which uses only information in the dictionary definitions, in contrast with previous work on English which uses hierarchical information in addition to dictionary definitions. We extend the definition based semantic similarity measure with distributional similarity applied to the words in different definitions. This increases the recall of our method and in some cases, precision as well.

1 Introduction

Word sense disambiguation (WSD) has been an active area of research over the last decade because many researches believe it will be important for applications which require, or would benefit from, some degree of semantic interpretation. There has been considerable skepticism over whether WSD will actually improve performance of applications, but we are now starting to see improvement in performance due to WSD in cross-lingual information retrieval (Clough and Stevenson, 2004; Vossen et al., 2006) and machine translation (Carpuat and Wu, 2007; Chan et al., 2007) and we hope that other applications such as question-answering, text simplification and summarisation might also benefit as WSD methods improve.

In addition to contextual evidence, most WSD systems exploit information on the most likely meaning of a word regardless of context. This is a powerful back-off strategy given the skewed nature of word sense distributions. For example, in the English coarse grained all words task (Navigli et al., 2007) at the recent SemEval Workshop the baseline of choosing the most frequent sense using the first WordNet sense attained precision and recall of 78.9% which is only a few percent lower than the top scoring system which obtained 82.5%. This finding is in line with previous results (Snyder and Palmer, 2004). Systems using a first sense heuristic have relied on sense-tagged data or lexicographer judgment as to which is the predominant sense of a word. However sense-tagged data is expensive and furthermore the predominant sense of a word will vary depending on the domain (Koeling et al., 2005; Chan and Ng, 2007).

One direction of research following McCarthy et al. (2004) has been to learn the most predominant sense of a word automatically. McCarthy et al's method relies on two methods of similarity. Firstly, distributional similarity is used to estimate the predominance of a sense from the number of distributionally similar words and the strength of their distributional similarity to the target word. This is done on the premise that more prevalent meanings have more evidence in the corpus data used for the distributional similarity calculations and the distributionally similar words (nearest neighbours) to a target reflect the more predominant meanings as a consequence. Secondly, the senses in the sense inventory are linked to the nearest neighbours using semantic similarity which incorporates information from the sense inventory. It is this semantic similarity measure which is the focus of our paper in the context of the method for acquiring predominant senses.

Whilst the McCarthy et al.'s method works well for English, other inventories do not always have WordNet style resources to tie the nearest neighbours to the sense inventory. WordNet has many semantic relations as well as glosses associated with its synsets (near synonym sets). While traditional dictionaries do not organise senses into synsets, they do typically have sense definitions associated with the senses. McCarthy et al. (2004) suggest that dictionary definitions can be used with their method, however in the implementation of the measure based on dictionary definitions that they use, the dictionary definitions are extended to those of related words using the hierarchical structure of WordNet (Banerjee and Pedersen, 2002). This extension to the original method (Lesk, 1986) was proposed because there is not always sufficient overlap of the individual words for which semantic similarity is being computed. In this paper we refer to the original method (Lesk, 1986) as **lesk** and the extended measure proposed by Banerjee and Pedersen as **Elesk**.

This paper investigates the potential of using the overlap of dictionary definitions with the McCarthy et al.'s method. We test the method for obtaining a first sense heuristic using two publicly available datasets of sense-tagged data in Japanese, EDR (NICT, 2002) and the SENSEVAL-2 Japanese dictionary task (Shirai, 2001). We contrast an implementation of **lesk** (Lesk, 1986) which uses only dictionary definitions with the Jiang-Conrath measure (**jcn**) (Jiang and Conrath, 1997) which uses manually produced hyponym links and was used previously for this purpose on English datasets (McCarthy et al., 2004). The **jcn** measure is only applicable to the EDR dataset because the dictionary has hyponymy links which are not available in the SENSEVAL-2 Japanese dictionary task. We also propose a new extension to **lesk** which does not require hand-crafted hyponym links but instead uses distributional similarity to increase the possibilities for overlap of the word definitions. We refer to this new measure as **DSlesk**. We compare this to the original **lesk** on both datasets and show that it increases recall, and sometimes precision too whilst not requiring hyponym links.

In the next section we place our contribution in relation to previous work. In section 3 we summarise the methods we adopt from previous work, and describe our proposal for a semantic similarity method that can supplement the information from dictionary definitions with information from raw text. In section 4 we describe the experiments on EDR and the SENSEVAL-2 Japanese dictionary task and we conclude in section 5.

2 Related Work

This work builds upon that of McCarthy et al. (2004) which acquires predominant senses for target words from a large sample of text using distributional similarity (Lin, 1998) to provide evidence for predominance. The evidence from the distributional similarity is allocated to the senses using semantic similarity from WordNet (Patwardhan and Pedersen, 2003). We will describe the method more fully below in section 3. McCarthy et al. (2004) reported results for English using their automatically acquired first sense heuristic on SemCor (Miller et al., 1993) and the SENSEVAL-2 English all words dataset (Snyder and Palmer, 2004). The results from this are promising, given that hand-labelled data is not required. On polysemous nouns from SemCor they obtained 48% WSD using their method with **Elesk** and 46% with **jcn** where the random baseline was 24% and the upper-bound was 67% (derived from the SemCor test data itself). On SENSEVAL-2 all words dataset using the **jcn** measure [1] they obtained 63% recall which is encouraging compared to the

[1] They did not apply **lesk** to this dataset.

SemCor heuristic which obtained 68% but requires hand-labelled data. The upper-bound on the dataset was 72% from the test data itself. These results crucially depend on the information in the sense inventory WordNet. WordNet contains hierarchical relations between word senses which are used in both **jcn** and **Elesk**. There is an issue that such information may not be available in other sense inventories, and other inventories will be needed for other languages. In this paper, we implement the **lesk** semantic similarity (Lesk, 1986) for the two Japanese lexicons used in our test datasets, i) the EDR dictionary (NICT, 2002) ii) the Iwanami Kokugo Jiten Dictionary (Nishio et al., 1994). We investigate the potential of **lesk** and **jcn**, where the latter is applicable. In addition to implementing the original **lesk** measure, we propose an extension to the method inspired by Mihalcea et al. (2006). Mihalcea et al. (2006) used various text based similarity measures, including WordNet and corpus based similarity methods, to determine if two phrases are paraphrases. They contrasted this approach with previous methods which used overlap of the words between the candidate paraphrases. For each word in each of the two texts they obtain the maximum similarity between the word and any of the words from the putative paraphrase. The similarity scores for each word of both phrases contribute to an overall semantic similarity between 0 and 1 and a threshold of 0.5 is used to decide if the candidate phrases are paraphrases. In our work, we compare glosses of words senses (senses of the target word and senses of the nearest neighbour) rather than paraphrases. In this approach we extend the definition overlap by considering the distributional similarity (Lin, 1998) rather than identify of the words in the two definitions.

In addition to McCarthy et al. (2004) there are other approaches to finding predominant senses. Chan and Ng (2005) use parallel data to provide estimates for sense frequency distributions to feed into a supervised WSD system. Mohammad and Hirst (2006) propose an approach to acquiring predominant senses from corpora which makes use of the category information in the Macquarie Thesaurus (Barnard, 1986). Lexical chains (Galley and McKeown, 2003) may also provide a useful first sense heuristic (Brody et al., 2006) but are produced using WordNet relations. We use the McCarthy et al. approach because this is applicable without aligned corpus data, semantic category and relation information and is applicable to any language assuming the minimum requirements of i) dictionary definitions associated with the sense inventory and ii) raw corpus data. We adapt their technique to remove the reliance on hyponym links.

3 Gloss-based semantic similarity

We first summarise the McCarthy et al. method and the WordNet based semantic similarity functions (**jcn** and **Elesk**) that they use for automatic acquisition of a first sense heuristic applied to disambiguation of English WordNet datasets. We then describe the additional semantic similarity method that we propose for comparison with **lesk** and **jcn**.

McCarthy et al. use a distributional similarity thesaurus acquired from corpus data using the method of Lin (1998) for finding the predominant sense of a word where the senses are defined by WordNet. The thesaurus provides the k nearest neighbours to each target word, along with the distributional similarity score between the target word and its neighbour. The WordNet similarity package (Patwardhan and Pedersen, 2003) is used to weight the contribution that each neighbour makes to the various senses of the target word.

Let w be a target word and $N_w = \{n_1, n_2 ... n_k\}$ be the ordered set of the top scoring k neighbours of w from the thesaurus with associated distributional similarity scores $\{dss(w,n_1), dss(w,n_2), ... dss(w,n_k)\}$ using (Lin, 1998). Let $senses(w)$ be the set of senses of w for each sense of w ($ws_i \in senses(w)$) a ranking is obtained using:

$Prevalence\ Score(ws_i) =$

$$\sum_{n_j \in N_w} dss(w,n_j) \times \frac{wnss(ws_i,n_j)}{\sum_{ws_{i'} \in senses(w)} wnss(ws_{i'},n_j)} \quad (1)$$

where $wnss$ is the maximum WordNet similarity score between ws_i and the WordNet sense of the neighbour (n_j) that maximises this score. McCarthy et al. compare two different WordNet similarity scores, **jcn** and **Elesk**.

jcn (Jiang and Conrath, 1997) uses corpus data to estimate a frequency distribution over the classes (synsets) in the WordNet hierarchy. Each synset, is incremented with the frequency counts from the corpus of all words belonging to that synset, directly or via the hyponymy relation. The frequency data is used to calculate the "information content" (IC) of a class or sense (s):

$$IC(s) = -log(p(s))$$

Jiang and Conrath specify a distance measure between two senses ($s1, s2$):

$$D_{jcn}(s1,s2) = IC(s1) + IC(s2) - 2 \times IC(s3)$$

where the third class ($s3$) is the most informative, or most specific, superordinate synset of the two senses $s1$ and $s2$. This is transformed from a distance measure in the WordNet Similarity package by taking the reciprocal:

$$jcn(s1,s2) = 1/D_{jcn}(s1,s2)$$

McCarthy et al. use the above measure with ws_i as $s1$ and whichever sense of the neigbour (n_j) that maximises this WordNet similarity score.

Elesk (Banerjee and Pedersen, 2002) extends the original **lesk** algorithm (Lesk, 1986) so we describe that original algorithm **lesk** first. This simply calculates the overlap of the content words in the definitions, frequently referred to as glosses, of the two word senses.

$$lesk(s1,s2) = \sum_{a \in g_1} member(a, g_2)$$

$$member(a, g_2) = \begin{cases} 1 & \text{if } a \text{ appears in } g_2 \\ 0 & otherwise \end{cases}$$

where g_1 is the gloss of word sense $s1$, g_2 is the gloss of $s2$ and a is one of words appearing in g_1. In **Elesk** which McCarthy et al. use the measure is extended by considering related synsets to $s1$ and $s2$, again where $s1$ is ws_i and $s2$ is the sense from all senses of n_j that maximises the **Elesk** WordNet similarity score. **Elesk** relies heavily on the relationships that are encoded in WordNet such as hyponymy and meronymy. Not all languages have resources supplied with these relations, and where they are supplied there may not be as much detail as there is in WordNet.

In this paper we will examine the use of **jcn** and the original **lesk** in Japanese on the EDR dataset to see how well the pure definition based measure fares compared to one using hyponym links. EDR has hyponym links so we can make this comparison. The performance of **jcn** will depend on the coverage of the hyponym links. For **lesk** meanwhile there is an issue that using only overlap of sense definitions may give poor results because the sense definitions are usually succinct and the overlap of words may be low. For example, given the glosses for the words *pigeon* and *bird*:[2]

> *pigeon: a fat grey and white bird with short legs.*
> *bird: a creature that is covered with feathers and has wings and two legs.*

If only content words are considered then there is only one word (*leg*) which overlaps in the two glosses, so the resultant **lesk** score is low (1) even though the word *pigeon* is intuitively similar to *bird*.

The **Elesk** extension addressed this issue using WordNet relations to extend the definitions over which the overlap is calculated for a given pair of senses. We propose addressing the same issue using corpus data to supplement the **lesk** overlap measure. We propose using distributional similarity (using (Lin, 1998)) as an approximation of semantic distance between the words in the two glosses, rather than requiring an exact match. We refer to this measure as **DSlesk** as defined:

$$DSlesk(s1,s2) = \frac{1}{|a \in g_1|} \sum_{a \in g_1} \max_{b \in g_2} dss(a,b) \quad (2)$$

where g_1 is the gloss of word sense $s1$, g_2 is the gloss of $s2$, again $s1$ is the target word sense ws_i in equation 1 for which we are obtaining the predominance ranking score and $s2$ is whichever sense of the neighbour (n_j) in equation 1 which maximises this semantic similarity score, as McCarthy et al. did with the *wnss* in equation 1. a (b) is a word appearing in g_1 (g_2).

In the calculation of equation (2), we first extract the most similar word b from g_2 to each word (a) in

[2] These two glosses are defined in OXFORD Advanced Learner's Dictionary.

$dss(bird, creature) = 0.84$, $dss(bird, feather) = 0.77$,
$dss(bird, wing) = 0.55$, $dss(bird, leg) = 0.43$,
$dss(leg, creature) = 0.56$, $dss(leg, feather) = 0.66$,
$dss(leg, wing) = 0.74$, $dss(leg, leg) = 1.00$

Figure 1: Examples of distributional similarity

the gloss of $s1$. We then output the average of the maximum distributional similarity of all the words in g_1 to any of the words in g_2 as the similarity score between $s1$ and $s2$. We acknowledge that **DSlesk** is not symmetrical since it depends on the number of words in the gloss of $s1$, but not $s2$. Also our summation is over these words in $s1$ and we are not looking for identity but maximum distributional similarity with any of the words in g_2 so the summation will not give the same result as if we did the summation over the words in g_2. It is perfectly reasonable to have a semantic similarity measure which is not symmetrical. One may want a measure where a more specific sense, such as the meat sense of *chicken* is closer to the "animal flesh used as food" sense of meat than vice versa. We do not believe that this asymmetry is problematic for our application as all the senses of w which we are ranking are all treated equally with respect to the neighbour n, and the ranking measure is concerned with finding evidence for the meaning of w, which we do by focusing on its definitions, and not the meaning of n. It would however be worthwhile investigating symmetrical versions of the score in the future.

Here is an example given the definitions of *bird* and *pigeon* above and the distributional similarity scores of all combinations of the two nouns as shown in Figure 1. In this case, the similarity is estimated as $1/2(0.84 + 1.00) = 0.92$.

4 Experiments

To investigate how well the McCarthy et al. method ports to other language, we conduct empirical evaluation of word sense disambiguation by using the two available sense-tagged datasets, EDR and the SENSEVAL-2 Japanese dictionary task. In the experiments, we compare the three semantic similarities, **jcn**, **lesk** and **DSlesk**[3], for use in the method to find the most likely sense in the set of word senses defined in each inventory following the approach of McCarthy et al. (2004). For the thesaurus construction we used <verb, case, noun> triplets extracted from Japanese newspaper articles (9 years of the Mainichi Shinbun (1991-1999) and 10 years of the Nihon Keizai Shinbun (1991-2000)) and parsed by CaboCha (Kudo and Matsumoto, 2002). This resulted in 53 million triplet instances for acquiring the distributional thesaurus. We adopt the similarity score proposed by Lin (1998) as the distributional similarity score and use 50 nearest neighbours in line with McCarthy et al.

For the random baseline we select one word sense at random for each word token and average the precision over 100 trials. For contrast with a supervised approach we show the performance if we use hand-labelled training data for obtaining the predominant sense of the test words. This method usually outperforms an automatic approach, but crucially relies on there being hand-labelled data which is expensive to produce. The method cannot be applied where there is no hand-labelled training data, it will be unreliable for low frequency data and a general dataset may not be applicable when one moves to domain specific text (Koeling et al., 2005). Since we are not using context for disambiguation, but just a first sense heuristic, we also give the upper-bound which is the first sense heuristic calculated from the test data itself.

4.1 EDR

We conduct empirical evaluation using 3,836 polysemous nouns in the sense-tagged corpus provided with EDR (183,502 instances) where the glosses are defined in the EDR dictionary. We evaluated on this dataset using WSD precision and recall of this corpus using only our first-sense heuristic (no context). The results are shown in Table 1. The WSD performance of all the automatic methods is much lower than the supervised method, however, the main point of this paper is to compare the McCarthy et al. method for finding a first sense in Japanese using **jcn**, **lesk** and

[3] **Elesk** can be used when several semantic relations such as hypnoymy and meronomy are available. However, we cannot directly apply **Elesk** as it was used in (McCarthy et al., 2004) to our experiments because the meronomy relation is not defined in the EDR dictionary. In the experiments reported here we focus on the comparison of the three similarity measures **jcn**, **lesk** and **DSlesk** for use in the method to determine the predominant sense of each word. We leave further exploration of other adaptations of semantic similarity scores for future work.

Table 1: Results of EDR

	recall	precision
baseline	0.402	0.402
jcn	0.495	0.495
lesk	0.474	0.488
DSlesk	0.495	0.495
upper-bound	0.745	0.745
supervised	0.731	0.731

Table 2: Precision on EDR at low frequencies

	all	freq ≤ 10	freq ≤ 5
baseline	0.402	0.405	0.402
jcn	0.495	0.445	0.431
lesk	0.474	0.448	0.426
DSlesk	0.495	0.453	0.433
upper-bound	0.745	0.674	0.639
supervised	0.731	0.519	0.367

Table 3: Results of SENSEVAL-2

	precision = recall	
	fine	coarse
baseline	0.282	0.399
lesk	0.344	0.501
DSlesk	0.386	0.593
upper-bound	0.747	0.834
supervised	0.742	0.842

DSlesk. Table 1 shows that **DSlesk** is comparable to **jcn** without the requirement for semantic relations such as hyponymy.

Furthermore, we evaluate precision of each method at low frequencies of words (≤ 10, ≤ 5), shown in Table 2. Table 2 shows that all methods for finding a predominant sense outperform the supervised one for items with little data (≤ 5), indicating that these methods robustly work even for low frequency data where hand-tagged data is unreliable.

Whilst the results are significantly different to the baseline [4] we note that the difference to the random baseline is less than for McCarthy et al. who obtained 48% for **Elesk** on polysemous nouns in SemCor and 46% for **jcn** against a random baseline of 24%. These differences are probably explained by differences in the lexical resources. Both **Elesk** and **jcn** rely on semantic relations including hyponymy with **Elesk** also using the glosses. **jcn** in both approaches use the hyponym links. WordNet 1.6 (used by McCarthy et al.) has 66025 synsets with 66910 hyponym links between these [5]. For EDR there are 166868 nodes (word sense groupings) and 53747 hyponym links. So in EDR the ratio of these links to the nodes is much lower. This and other differences between EDR and WordNet are likely to be the reason for the difference in results.

4.2 SENSEVAL-2

We also evaluate the performance using the Japanese dictionary task in SENSEVAL-2 (Shirai, 2001). In this experiment, we use 50 nouns (5,000 instances). For this task, since semantic relations such as hyponym links are not defined, use of **jcn** is not possible. Therefore, we just compare **lesk** and **DSlesk** along with our random baseline, the supervised approach and the upper-bound as before.

The results are evaluated in two ways; one is for fine-grained senses in the original task definition and the other is coarse-grained version which is evaluated discarding the finer categorical information of each definition. The results are shown in Table 3. As with the EDR results, all unsupervised methods significantly outperform the baseline method, though the supervised methods still outperform the unsupervised ones. In this experiment, **DSlesk** is also significantly better than **lesk** in both fine and coarse-grained evaluations. It indicates that applying distributional similarity score to calculating inter-gloss similarities improves performance.

5 Conclusion

In this paper, we examined different measures of semantic similarity for finding a first sense heuristic for WSD automatically in Japanese. We defined a new gloss-based similarity (**DSlesk**) and evaluated the performance on two Japanese WSD datasets, outperforming **lesk** and achieving a performance comparable to the **jcn** method which relies on hyponym links which are not always available.

[4] For significance testing we used McNemar's test $\alpha = 0.05$.

[5] These figures are taken from http://www.lsi.upc.es/~batalla/wnstats.html#wn16

There are several issues for future directions of automatic detection of a first sense heuristic. In this paper, we proposed an adaptation of the **lesk** measure of gloss-based similarity, by using the average similarity between nouns in the two glosses under comparison in a bag-of-words approach without recourse to other information. However, it would be worthwhile exploring other information in the glosses, such as words of other PoS and predicate argument relations. We also hope to investigate applying alignment techniques introduced for entailment recognition (Hickl and Bensley, 2007).

Another important issue in WSD is to group fine-grained word senses into clusters, making the task suitable for NLP applications (Ide and Wilks, 2006). We believe that our gloss-based similarity **DSlesk** might be very suitable for this task and we plan to investigate the possibility.

There are other approaches we would like to explore in future. Mihalcea (2005) uses dictionary definitions alongside graphical algorithms for unsupervised WSD. Whilst the results are not directly comparable to ours because we have not included contextual evidence in our models, it would be worthwhile exploring if unsupervised graphical models using only the definitions we have in our lexical resources can perform WSD on a document and give more reliable first sense heuristics.

Acknowledgements

This work was supported by the UK EPSRC project EP/C537262 'Ranking Word Senses for Disambiguation: Models and Applications', and a UK Royal Society Dorothy Hodgkin Fellowship to the second author. We would like to thank John Carroll for several useful discussions on this work.

References

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted Lesk algorithm for word sense disambiguation using WordNet. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-02)*, Mexico City.

J.R.L. Barnard, editor. 1986. *Macquaire Thesaurus*. Macquaire Library, Sydney.

Samuel Brody, Roberto Navigli, and Mirella Lapata. 2006. Ensemble methods for unsupervised wsd. 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.

Marine Carpuat and Dekai Wu. 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)*, pages 61–72, Prague, Czech Republic, June. Association for Computational Linguistics.

Yee Seng Chan and Hwee Tou Ng. 2005. Word sense disambiguation with distribution estimation. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI 2005)*, Edinburgh, Scotland.

Yee Seng Chan and Hwee Tou Ng. 2007. Domain adaptation with active learning for word sense disambiguation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic, June. Association for Computational Linguistics.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic, June. Association for Computational Linguistics.

Paul Clough and Mark Stevenson. 2004. Evaluating the contribution of EuroWordNet and word sense disambiguation to cross-language retrieval. In *Second International Global WordNet Conference (GWC-2004)*, pages 97–105.

Michel Galley and Kathleen McKeown. 2003. Improving word sense disambiguation in lexical chaining. In *IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, Mexico, August 9-15, 2003*, pages 1486–1488. Morgan Kaufmann.

Andrew Hickl and Jeremy Bensley. 2007. A discourse commitment-based framework for recognizing textual entailment. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 171–176.

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.

Jay Jiang and David Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *International Conference on Research in Computational Linguistics*, Taiwan.

Rob Koeling, Diana McCarthy, and John Carroll. 2005. Domain-specific sense distributions and predominant sense acquisition. In *Proceedings of the joint conference on Human Language Technology and Empirical methods in Natural Language Processing*, pages 419–426, Vancouver, B.C., Canada.

Taku Kudo and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In *Proceedings of the 6th Conference on Natural Language Learning 2002* (*CoNLL*), pages 63–69.

M. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from and ice cream cone. In *Proceedings of the ACM SIGDOC Conference*, pages 24–26, Toronto, Canada.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL 98*, Montreal, Canada.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 280–287, Barcelona, Spain.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the American Association for Artificial Intelligence (AAAI 2006)*, Boston, MA, July.

Rada Mihalcea. 2005. Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In *Proceedings of the joint conference on Human Language Technology and Empirical methods in Natural Language Processing*, Vancouver, B.C., Canada.

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

Saif Mohammad and Graeme Hirst. 2006. Determining word sense dominance using a thesaurus. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, pages 121–128, Trento, Italy, April.

Roberto Navigli, C. Litkowski, Kenneth, and Orin Hargraves. 2007. SemEval-2007 task 7: Coarse-grained English all-words task. In *Proceedings of ACL/SIGLEX SemEval-2007*, pages 30–35, Prague, Czech Republic.

NICT. 2002. EDR electronic dictionary version 2.0, technical guide. http://www2.nict.go.jp/kk/e416/EDR/.

Minoru Nishio, Etsutaro Iwabuchi, and Shizuo Mitzutani. 1994. Iwanami kokugo jiten dai go han.

Siddharth Patwardhan and Ted Pedersen. 2003. The CPAN WordNet::Similarity Package. http://search.cpan.org/author/SID/WordNet-Similarity-0.03/.

Kiyoaki Shirai. 2001. SENSEVAL-2 Japanese Dictionary Task. In *Proceedings of the* SENSEVAL-2 *workshop*, pages 33–36.

Benjamin Snyder and Martha Palmer. 2004. The English all-words task. In *Proceedings of the ACL* SENSEVAL-3 *workshop*, pages 41–43, Barcelona, Spain.

Piek Vossen, German Rigau, Inaki Alegria, Eneko Agirre, David Farwell, and Manuel Fuentes. 2006. Meaningful results for information retrieval in the meaning project. In *Proceedings of the 3rd Global WordNet Conference*. http://nlpweb.kaist.ac.kr/gwc/.