Applying a Naive Bayes Similarity Measure to Word Sense Disambiguation

Tong Wang
University of Toronto
tong@cs.toronto.edu

Graeme Hirst
University of Toronto
gh@cs.toronto.edu

Abstract

We replace the overlap mechanism of the Lesk algorithm with a simple, general-purpose Naive Bayes model that measures *many-to-many* association between two sets of random variables. Even with simple probability estimates such as maximum likelihood, the model gains significant improvement over the Lesk algorithm on word sense disambiguation tasks. With additional lexical knowledge from WordNet, performance is further improved to surpass the state-of-the-art results.

1 Introduction

To disambiguate a homonymous word in a given context, Lesk (1986) proposed a method that measured the degree of overlap between the glosses of the target and context words. Known as the Lesk algorithm, this simple and intuitive method has since been extensively cited and extended in the word sense disambiguation (WSD) community. Nonetheless, its performance in several WSD benchmarks is less than satisfactory (Kilgarriff and Rosenzweig, 2000; Vasilescu et al., 2004). Among the popular explanations is a key limitation of the algorithm, that "Lesk's approach is very sensitive to the exact wording of definitions, so the absence of a certain word can radically change the results." (Navigli, 2009).

Compounding this problem is the fact that many Lesk variants limited the concept of overlap to the literal interpretation of string matching (with their own variants such as length-sensitive matching (Banerjee and Pedersen, 2002), etc.), and it was not until recently that overlap started to take on other forms such as tree-matching (Chen et al., 2009) and vector space models (Abdalgader and Skabar, 2012; Raviv et al., 2012; Patwardhan and Pedersen, 2006). To address this limitation, a Naive Bayes model (NBM) is proposed in this study as a novel, probabilistic treatment of overlap in gloss-based WSD.

2 Related Work

In the extraordinarily rich literature on WSD, we focus our review on those closest to the topic of Lesk and NBM. In particular, we opt for the "simplified Lesk" (Kilgarriff and Rosenzweig, 2000), where inventory senses are assessed by gloss-context overlap rather than gloss-gloss overlap. This particular variant prevents proliferation of gloss comparison on larger contexts (Mihalcea et al., 2004) and is shown to outperform the original Lesk algorithm (Vasilescu et al., 2004).

To the best of our knowledge, NBMs have been employed exclusively as classifiers in WSD — that is, in contrast to their use as a similarity measure in this study. Gale et al. (1992) used NB classifier resembling an information retrieval system: a WSD instance is regarded as a document d, and candidate senses are scored in terms of "relevance" to d. When evaluated on a WSD benchmark (Vasilescu et al., 2004), the algorithm compared favourably to Lesk variants (as expected for a supervised method). Pedersen (2000) proposed an ensemble model with multiple NB classifiers differing by context window size. Hristea (2009) trained an unsupervised NB classifier using the EM algorithm and empirically demonstrated the benefits of WordNet-assisted (Fellbaum, 1998) feature selection over local syntactic features.

Among Lesk variants, Banerjee and Pedersen (2002) extended the gloss of both inventory senses and the context words to include words in their related synsets in WordNet. Senses were scored by the sum of overlaps across all relation pairs, and the effect of individual relation pairs was evaluated in a later work (Banerjee and Pedersen, 2003). Overlap was assessed by string matching, with the number of matching words squared so as to assign

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 531–537,
Baltimore, Maryland, USA, June 23-25 2014.

higher scores to multi-word overlaps.

Breaking away from string matching, Wilks et al. (1990) measured overlap as similarity between gloss- and context-vectors, which were aggregated word vectors encoding second order co-occurrence information in glosses. An extension by Patwardhan and Pedersen (2006) differentiated context word senses and extended shorter glosses with related glosses in WordNet. Patwardhan et al. (2003) measured overlap by *concept similarity* (Budanitsky and Hirst, 2006) between each inventory sense and the context words. Gloss overlaps from their earlier work actually out-performed all five similarity-based methods.

More recently, Chen et al. (2009) proposed a tree-matching algorithm that measured gloss-context overlap as the weighted sum of dependency-induced lexical distance. Abdalgader and Skabar (2012) constructed a *sentential* similarity measure (Li et al., 2006) using *lexical* similarity measures (Budanitsky and Hirst, 2006), and overlap was measured by the cosine of their respective sentential vectors. A related approach (Raviv et al., 2012) also used Wikipedia-induced concepts to encoded sentential vectors. These systems compared favourably to existing methods in WSD performance, although by using sense frequency information, they are essentially supervised methods.

Distributional methods have been used in many WSD systems in quite different flavours than the current study. Kilgarriff and Rosenzweig (2000) proposed a Lesk variant where each gloss word is weighted by its *idf* score in relation to all glosses, and gloss-context association was incremented by these weights rather than binary, overlap counts. Miller et al. (2012) used distributional thesauri as a knowledge base to increase overlaps, which were, again, assessed by string matching.

In conclusion, the majority of Lesk variants focused on extending the gloss to increase the chance of overlapping, while the proposed NBM aims to make better use of the limited lexical knowledge available. In contrast to string matching, the probabilistic nature of our model offers a "softer" measurement of gloss-context association, resulting in a novel approach to unsupervised WSD with state-of-the-art performance in more than one WSD benchmark (Section 4).

3 Model and Task Descriptions

3.1 The Naive Bayes Model

Formally, given two sets $\mathbf{e} = \{e_i\}$ and $\mathbf{f} = \{f_j\}$ each consisting of multiple random events, the proposed model measures the probabilistic association $p(\mathbf{f}|\mathbf{e})$ between $\mathbf{e}$ and $\mathbf{f}$. Under the assumption of conditional independence among the events in each set, a Naive Bayes treatment of the measure can be formulated as:

$$p(\mathbf{f}|\mathbf{e}) = \prod_j p(f_j|\{e_i\}) = \prod_j \frac{p(\{e_i\}|f_j)p(f_j)}{p(\{e_i\})} = \frac{\prod_j [p(f_j)\prod_i p(e_i|f_j)]}{\prod_j \prod_i p(e_i)}, \quad (1)$$

In the second expression, Bayes's rule is applied not only to take advantage of the conditional independence among e_i's, but also to facilitate probability estimation, since $p(\{e_i\}|f_j)$ is easier to estimate in the context of WSD, where sample spaces of $\mathbf{e}$ and $\mathbf{f}$ become asymmetric (Section 3.2).

3.2 Model Application in WSD

In the context of WSD, $\mathbf{e}$ can be regarded as an instance of a polysemous word w, while $\mathbf{f}$ represents certain lexical knowledge about the sense s of w manifested by $\mathbf{e}$.[1] WSD is thus formulated as identifying the sense s^* in the sense inventory $\mathscr{S}$ of w s.t.:

$$s^* = \arg\max_{s \in \mathscr{S}} p(\mathbf{f}|\mathbf{e}) \quad (2)$$

In one of their simplest forms, e_i's correspond to co-occurring words in the instance of w, and f_j's consist of the gloss words of sense s. Consequently, $p(\mathbf{f}|\mathbf{e})$ is essentially measuring the association between context words of w and definition texts of s, i.e., the gloss-context association in the simplified Lesk algorithm (Kilgarriff and Rosenzweig, 2000). A major difference, however, is that instead of using hard, overlap counts between the two sets of words from the gloss and the context, this probabilistic treatment can implicitly model the distributional similarity among the elements e_i and f_j (and consequently between the sets $\mathbf{e}$ and $\mathbf{f}$) over a wider range of contexts. The result is a "softer" proxy of association than the binary view of overlaps in existing Lesk variants.

The foregoing discussion offers a second motivation for applying Bayes's rule on the second

[1] Think of the notations $\mathbf{e}$ and $\mathbf{f}$ mnemonically as *exemplars* and *features*, respectively.

Senses	Hypernyms	Hyponyms	Synonyms
factory	building complex, complex	brewery, factory, mill, ...	works, industrial plant
life form	organism, being	perennial, crop...	flora, plant life

Table 1: Lexical knowledge for the word *plant* under its two meanings *factory* and *life form*.

expression in Equation (1): it is easier to estimate $p(e_i|f_j)$ than $p(f_j|e_i)$, since the vocabulary for the lexical knowledge features (f_j) is usually more limited than that of the contexts (e_i) and hence estimation of the former suffices on a smaller amount of data than that of the latter.

3.3 Incorporating Additional Lexical Knowledge

The input of the proposed NBM is bags of words, and thus it is straightforward to incorporate various forms of lexical knowledge (LK) for word senses: by concatenating a tokenized knowledge source to the existing knowledge representation **f**, while the similarity measure remains unchanged.

The availability of LK largely depends on the sense inventory used in a WSD task. WordNet senses are often used in Senseval and SemEval tasks, and hence senses (or synsets, and possibly their corresponding word forms) that are semantic related to the inventory senses under WordNet relations are easily obtainable and have been exploited by many existing studies.

As pointed out by Patwardhan et al. (2003), however, "not all of these relations are equally helpful." Relation pairs involving hyponyms were shown to result in better F-measure when used in gloss overlaps (Banerjee and Pedersen, 2003). The authors attributed the phenomenon to the the multitude of hyponyms compared to other relations. We further hypothesize that, beyond sheer numbers, synonyms and hyponyms offer stronger semantic specification that helps distinguish the senses of a given ambiguous word, and thus are more effective knowledge sources for WSD.

Take the word *plant* for example. Selected hypernyms, hyponyms, and synonyms pertaining to its two senses *factory* and *life form* are listed in Table 1. Hypernyms can be overly general terms (e.g., *being*). Although conceptually helpful for humans in coarse-grained WSD, this generality is likely to inflate the hypernyms' probabilistic estimation. Hyponyms, on the other hand, help specify their corresponding senses with information that is possibly missing from the often overly brief glosses: the many technical terms as hyponyms in Table 1 — though rare — are likely to occur in the (possibly domain-specific) contexts that are highly typical of the corresponding senses. Particularly for the NBM, the co-occurrence is likely to result in stronger gloss-definition associations when similar contexts appear in a WSD instance.

We also observe that some semantically related words appear under rare senses (e.g., *still* as an alcohol-manufacturing plant, and *annual* as a one-year-life-cycle plant; omitted from Table 1). This is a general phenomenon in gloss-based WSD and is beyond the scope of the current discussion.[2] Overall, all three sources of LK may complement each other in WSD tasks, with hyponyms particularly promising in both quantity and quality compared to hypernyms and synonyms.[3]

3.4 Probability Estimation

A most open-ended question is how to estimate the probabilities in Equation (1). In WSD in particular, the estimation concerns the marginal and conditional probabilities of and between word tokens. Many options are available to this end in statistical machine learning (MLE, MAP, etc.), information theory (Church and Hanks, 1990; Turney, 2001), as well as the rich body of research in lexical semantic similarity Resnik, 1995; Jiang and Conrath, 1997; Budanitsky and Hirst, 2006).

Here we choose maximum likelihood — not only for its simplicity, but also to demonstrate model strength with a relatively crude probability estimation. To avoid underflow, Equation (1) is estimated as the following log probability:

$$\begin{aligned}&\sum_i \log \frac{c(f_j)}{c(\cdot)} + \sum_i \sum_j \log \frac{c(e_i, f_j)}{c(f_j)} - |\mathbf{f}| \sum_j \log \frac{c(e_i)}{c(\cdot)} \\ =&(1-|\mathbf{e}|) \sum_i \log c(f_j) - |\mathbf{f}| \sum_j \log c(e_i) \\ &+ \sum_i \sum_j \log c(e_i, f_j) + |\mathbf{f}|(|\mathbf{e}|-1) \log c(\cdot),\end{aligned}$$

where $c(x)$ is the count of word x, $c(\cdot)$ is the corpus

[2] We do, however, refer curious readers to the work of Raviv et al. (2012) for a novel treatment of a similar problem.

[3] Note that LK expansion is a feature of our model rather than a requirement. What type of knowledge to include is eventually a decision made by the user based on the application and LK availability.

size, $c(x,y)$ is the joint count of x and y, and $|\mathbf{v}|$ is the dimension of vector $\mathbf{v}$.

Nonetheless, we do investigate how model performance responds to estimation quality. Specifically in WSD, a *source corpus* is defined as the source of the majority of the WSD instances in a given dataset, and a *baseline corpus* of a smaller size and less resemblance to the instances is used for all datasets. The assumption is that a source corpus offers better estimates for the model than the baseline corpus, and difference in model performance is expected when using probability estimation of different quality.

4 Evaluation

4.1 Data, Scoring, and Pre-processing

Various aspects of the model discussed in Section 3 are evaluated in the English lexical sample tasks from Senseval-2 (Edmonds and Cotton, 2001) and SemEval-2007 (Pradhan et al., 2007). Training sections are used as development data and test sections held out for final testing. Model performance is evaluated in terms of WSD accuracy using Equation (2) as the scoring function. Accuracy is defined as the number of correct responses over the number of instances. Because it is a rare event for the NBM to produce identical scores,[4] the model always proposes a unique answer and accuracy is thus equivalent to F-score commonly used in existing reports.

Multiword expressions (MWEs) in the Senseval-2 sense inventory are not explicitly marked in the contexts. Several of the top-ranking systems implemented their own MWE detection algorithms (Kilgarriff and Rosenzweig, 2000; Litkowski, 2002). Without digressing to the details of MWE detection — and meanwhile, to ensure fair comparison with existing systems — we implement two variants of the prediction module, one completely ignorant of MWE and defaulting to `INCORRECT` for all MWE-related answers, while the other assuming perfect MWE detection and performing regular disambiguation algorithm on the MWE-related senses (*not* defaulting to `CORRECT`). All results reported for Senseval-2 below are harmonic means of the two outcomes.

Each inventory sense is represented by a set of LK tokens (e.g., definition texts, synonyms, etc.) from their corresponding WordNet synset (or in the coarse-grained case, a concatenation of tokens from all synsets in a sense group). The *MIT-JWI* library (Finlayson, 2014) is used for accessing WordNet. Usage examples in glosses (included by the library by default) are removed in our experiments.[5]

Basic pre-processing is performed on the contexts and the glosses, including lower-casing, stopword removal, lemmatization on both datasets, and tokenization on the Senseval-2 instances.[6] *Stanford CoreNLP*[7] is used for lemmatization and tokenization. Identical procedures are applied to all corpora used for probability estimation.

Binomial test is used for significance testing, and with one exception explicitly noted in Section 4.3, all differences presented are statistically highly significant ($p < 0.001$).

4.2 Comparing Lexical Knowledge Sources

To study the effect of different types of LK in WSD (Section 3.3), for each inventory sense, we choose synonyms (*syn*), hypernyms (*hpr*), and hyponyms (*hpo*) as extended LK in addition to its gloss. The WSD model is evaluated with gloss-only (*glo*), individual extended LK sources, and the combination of all four sources (*all*). The results are listed in Table 2 together with existing results (1st and 2nd correspond to the results of the top two unsupervised methods in each dataset).[8]

By using only glosses, the proposed model already shows statistically significant improvement over the basic Lesk algorithm (92.4% and 140.5% relative improvement in Senseval-2 coarse- and fine-grained tracks, respectively).[9] Moreover, comparison between coarse- and fine-grained tracks reveals interesting properties of different LK sources. Previous hypotheses (Section 3.3) are empirically confirmed that WSD perfor-

[4] This has never occurred in the hundreds of thousands of runs in our development process.

[5] We also compared the two Lesk baselines (with and without usage examples) on the development data but did not observe significant differences as reported by Kilgarriff and Rosenzweig (2000).

[6] The SemEval-2007 instances are already tokenized.

[7] `http://nlp.stanford.edu/software/corenlp.shtml`.

[8] We excluded the results of *UNED* (Fernández-Amorós et al., 2001) in Senseval-2 because, by using sense frequency information that is only obtainable from sense-annotated corpora, it is essentially a supervised system.

[9] Comparisons are made against the simplified Lesk algorithm (Kilgarriff and Rosenzweig, 2000) without usage examples. The comparison is unavailable in SemEval2007 since we have not found existing experiments with this exact configuration.

Dataset	*glo*	*syn*	*hpr*	*hpo*	*all*	1st	2nd	*Lesk*
Senseval-2 Coarse	.475	.478	.494	.518	**.523**	.469	.367	.262
Senseval-2 Fine	.362	.371	.326	.379	.388	**.412**	.293	.163
SemEval-2007	.494	.511	.507	.550	**.573**	.538	.521	–

Table 2: Lexical knowledge sources and WSD performance (*F-measure*) on the Senseval-2 (fine- and coarse-grained) and the SemEval-2007 dataset.

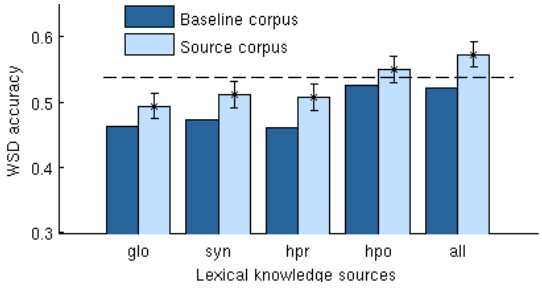

Figure 1: Model response to probability estimates of different quality on the SemEval-2007 dataset. Error bars indicate confidence intervals ($p < .001$), and the dashed line corresponds to the best reported result.

mance benefits most from hyponyms and least from hypernyms. Specifically, highly similar, fine-grained sense candidates apparently share more hypernyms in the fine-grained case than in the coarse-grained case; adding to the generality of hypernyms (both semantic and distributional), we postulate that their probability in the NBM is uniformly inflated among many sense candidates, and hence they decrease in distinguishability. Synonyms might help with regard to semantic specification, though their limited quantity also limits their benefits. These patterns on the LK types are consistent in all three experiments.

When including all four LK sources, our model outperforms the state-of-the-art systems with statistical significance in both coarse-grained tasks. For the fine-grained track, it achieves 2nd place after that of Tugwell and Kilgarriff (2001), which used a decision list (Yarowsky, 1995) on *manually selected* corpora evidence for each inventory sense, and thus is not subject to loss of distinguishability in the glosses as Lesk variants are.

4.3 Probability Estimation

To evaluate model response to probability estimation of different quality (Section 3.4), source corpora are chosen as the majority value of the *doc-source* attribute of instances in each dataset, namely, the *British National Corpus* for Senseval-2 (94%) and the *Wall Street Journal* for SemEval-2007 (86%). The *Brown Corpus* is shared by both datasets as the baseline corpus. Figure 1 shows the comparison on the SemEval-2007 dataset. Across all experiments, higher WSD accuracy is consistently witnessed using the source corpus; differences are statistically highly significant except for *hpo* (which is significant with $p < 0.01$).

5 Conclusions and Future Work

We have proposed a general-purpose Naive Bayes model for measuring association between two sets of random events. The model replaced string matching in the Lesk algorithm for word sense disambiguation with a probabilistic measure of gloss-context overlap. The base model on average more than doubled the accuracy of Lesk in Senseval-2 on both fine- and coarse-grained tracks. With additional lexical knowledge, the model also outperformed state of the art results with statistical significance on two coarse-grained WSD tasks.

For future work, we plan to apply the model in other shared tasks, including open-text WSD, so as to compare with more recent Lesk variants. We would also like to explore how to incorporate syntactic features and employ alternative statistical methods (e.g., parametric models) to improve probability estimation and inference. Other NLP problems involving compositionality in general might also benefit from the proposed many-to-many similarity measure.

Acknowledgments

This study is funded by the Natural Sciences and Engineering Research Council of Canada. We thank Afsaneh Fazly, Navdeep Jaitly, and Varada Kolhatkar for the many inspiring discussions, as well as the anonymous reviewers for their constructive advice.

References

Khaled Abdalgader and Andrew Skabar. Unsupervised similarity-based word sense disambiguation using context vectors and sentential word importance. *ACM Transactions on Speech and Language Processing*, 9(1):2:1–2:21, May 2012.

Satanjeev Banerjee and Ted Pedersen. An adapted Lesk algorithm for word sense disambiguation using WordNet. In *Computational Linguistics and Intelligent Text Processing*, pages 136–145. Springer, 2002.

Satanjeev Banerjee and Ted Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, volume 3, pages 805–810, 2003.

Alexander Budanitsky and Graeme Hirst. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006.

Ping Chen, Wei Ding, Chris Bowes, and David Brown. A fully unsupervised word sense disambiguation method using dependency knowledge. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 28–36, Stroudsburg, PA, USA, 2009.

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

Philip Edmonds and Scott Cotton. Senseval-2: Overview. In *Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5. Association for Computational Linguistics, 2001.

Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

David Fernández-Amorós, Julio Gonzalo, and Felisa Verdejo. The UNED systems at Senseval-2. In *The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 75–78. Association for Computational Linguistics, 2001.

Mark Alan Finlayson. Java libraries for accessing the Princeton WordNet: Comparison and evaluation. In *Proceedings of the 7th Global Wordnet Conference*, Tartu, Estonia, 2014.

William Gale, Kenneth Church, and David Yarowsky. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26(5-6):415–439, 1992.

Florentina Hristea. Recent advances concerning the usage of the Naïve Bayes model in unsupervised word sense disambiguation. *International Review on Computers & Software*, 4(1), 2009.

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

Adam Kilgarriff and Joseph Rosenzweig. Framework and results for English Senseval. *Computers and the Humanities*, 34(1-2):15–48, 2000.

Michael Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, pages 24–26, New York, New York, USA, 1986.

Yuhua Li, David McLean, Zuhair A Bandar, James D O'Shea, and Keeley Crockett. Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering*, 18(8):1138–1150, 2006.

Kenneth C. Litkowski. Sense information for disambiguation: Confluence of supervised and unsupervised methods. In *Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 47–53. Association for Computational Linguistics, July 2002.

Rada Mihalcea, Paul Tarau, and Elizabeth Figa. PageRank on semantic networks, with application to word sense disambiguation. In *Proceedings of the 20th International Conference on Computational Linguistics*, 2004.

Tristan Miller, Chris Biemann, Torsten Zesch, and Iryna Gurevych. Using distributional similarity for lexical expansion in knowledge-based word sense disambiguation. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 1781–1796, 2012.

Roberto Navigli. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):10:1–10:69, 2009.

Siddharth Patwardhan and Ted Pedersen. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. *Proceedings of the EACL 2006 Workshop Making Sense of Sense-Bringing Computational Linguistics and Psycholinguistics Together*, 1501:1–8, 2006.

Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. Using measures of semantic relatedness for word sense disambiguation. In *Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 241–257, 2003.

Ted Pedersen. A simple approach to building ensembles of Naive Bayesian classifiers for word sense disambiguation. In *Proceedings of the 1st Conference of North American Chapter of the Association for Computational Linguistics*, pages 63–69. Association for Computational Linguistics, 2000.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. SemEval-2007 task 17: English lexical sample, SRL and all words. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 87–92. Association for Computational Linguistics, 2007.

Ariel Raviv, Shaul Markovitch, and Sotirios-Efstathios Maneas. Concept-based approach to word sense disambiguation. In *Proceedings of the 26th Conference on Artificial Intelligence*, 2012.

Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, pages 448–453, San Francisco, CA, USA, 1995.

David Tugwell and Adam Kilgarriff. Wasp-bench: a lexicographic tool supporting word sense disambiguation. In *The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 151–154. Association for Computational Linguistics, 2001.

Peter Turney. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning*, pages 491–502, 2001.

Florentina Vasilescu, Philippe Langlais, and Guy Lapalme. Evaluating variants of the Lesk approach for disambiguating words. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, 2004.

Yorick Wilks, Dan Fass, Cheng-Ming Guo, James E. McDonald, Tony Plate, and Brian M. Slator. Providing machine tractable dictionary tools. *Machine Translation*, 5(2):99–154, 1990.

David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 189–196, 1995.