2002 IEEE International Symposium on Natural Language Processing
and Knowledge Engineering (NLPKE 2002). To appear.

Word Sense Disambiguation in Document Space

Krister Lindén†, Krista Lagus‡

†University of Helsinki, Department of General Linguistics
P.O.Box 9 (Siltavuorenpenger 20 A), FIN-00014 University of Helsinki, Finland.
E-mail: Krister.Linden@helsinki.fi
‡Helsinki University of Technology, Neural Networks Research Centre
P.O.Box 9800 (Tammasaarenkatu 3), FIN-02015 HUT, Finland
E-mail: Krista.Lagus@hut.fi.

Abstract— **We introduce a method for word sense disambiguation that uses an existing topical document map created with an unsupervised method (WEBSOM [1]) on a very large document collection. Results on the SENSEVAL−2 corpus indicate that the proposed method is statistically significantly better than the baselines and on a par with supervised methods.**

The method uses the document map as a representation of the semantic space of word contexts. The assumption is that similar meanings of a word have similar contexts, which are located in the same area on the self-organized document map. The results confirm this assumption.

The benefit of the proposed method is that a single general purpose representation of the semantic space can be used for all words and their word senses.

Keywords— **Word sense disambiguation, Self-organized document maps, Large corpora, Semantic space**

I. Introduction

WORD sense disambiguation is the task of automatically determining the appropriate senses of a word in context. It is an important and difficult problem with many practical consequences for language-technology applications in information retrieval, document classification, machine translation, spelling correction, parsing, speech synthesis as well as speech recognition. In order to create domain-independent applications in these areas word sense disambiguation is essential. So far many applications solve the problem by letting users manually switch between tailor-made domain-specific vocabularies. In contrast, successful word sense disambiguation could provide a seamless transition between domains.

The word sense disambiguation problem has been approached by traditional AI methods, such as hand-made rule sets or semantic networks, by knowledge-based methods using dictionaries or thesauri, and by corpus-based methods [2]. For a textbook introduction to word sense disambiguation, see [3]. The methods may vary in how different levels of context are selected and encoded. From a linguistic point of view the information included in the representation of context corresponds to approximations of morphological, syntactic and discourse context. Yarowsky [4] noted that there seems to be only one sense per collocation and that words tend to keep the same sense during a discourse. Later Martinez and Agirre [5] confirmed that the one-sense-per-collocation hypothesis holds, but it is weaker on fine-grained sense distinctions. In addition they showed that different topics tend to have different sets of collocations, which reduces the applicability of collocations over topic or genre variations. In this paper we evaluate genre and topic independent disambiguation effects. We therefore rely heavily on the one-sense-per-discourse hypothesis for single words. We take one document to be part of a discourse, but we assume no rigid division between discourses.

Corpus-based machine learning methods fall into two categories: unsupervised methods operate on untagged words in context, whereas supervised methods use sense-tagged words in context. For a recent comparison of algorithms, see [6], [7], and for some results of combining supervised and unsupervised methods, see e.g. [8]. In [9], an unsupervised method is proposed, which uses clusters of ambiguous words in context. The method achieves good results on a small set of words using only the two main senses of each word. The method created word clusters for the same genre and topics in the training data as in the test data.

We propose a hybrid technique, which uses a self-organizing map to create a representation of semantic space from a massive independent document collection and then calibrates the representation with a small batch of hand-tagged data from the same genre as the test data. However, the original training data for the self-organized document map was not selected for the set of words to be disambiguated and was from a different genre than the test data.

A mathematical structure of semantic space is discussed in [10]. Formally it is a quadruple $< A, B, S, M >$, where B is the set of basis elements, A is the mapping between particular basis elements and each word in the language, S is the similarity measure between vectors of basis elements, and M is a transformation between two semantic spaces, e.g. a dimensionality reduction. In [11], Steyvers and Tenenbaum show that large-scale natural language semantic structures such as thesauri are characterized by sparse connectivity and strong local clustering. In [12], Martinetz and Schulten show that self-organizing maps often are a local neighborhood preserving projection from a high dimensional space to a low dimensional display. This leads to an assumption that a document map created

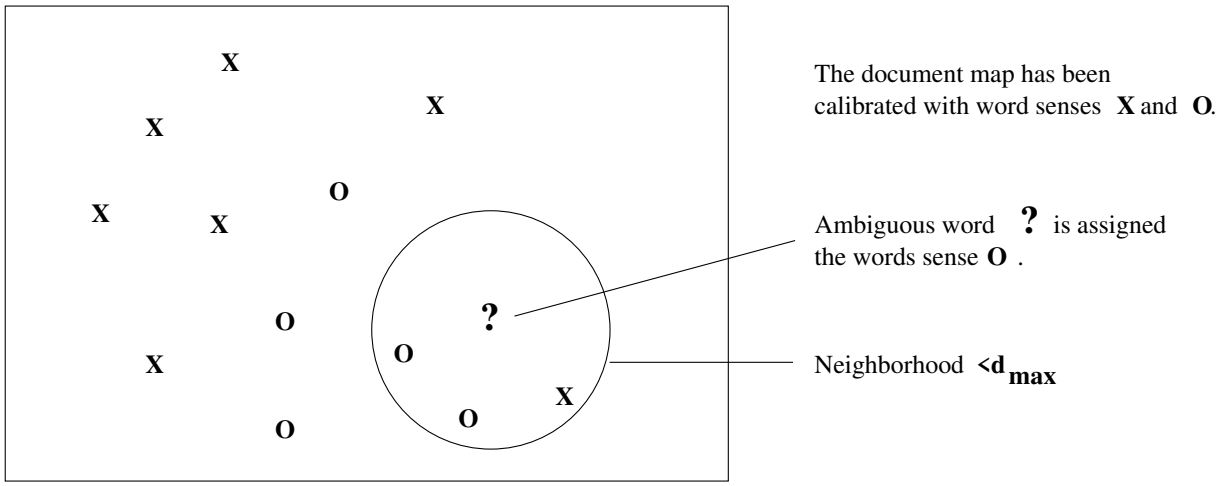

Fig. 1. Word sense disambiguation with a WEBSOM map

by the WEBSOM method [1] could serve as a locally accurate representation of a high dimensional semantic space, since it preserves the strong local clustering characteristic of large semantic structures. A self-organized document map represents the semantic space as ordered clusters of documents. The ordering thereby locally reflects the topical and discourse contexts of words.

In the same sense that a general language thesaurus is domain independent, we take the particular document map, the creation of which is described in [1], and evaluate our disambiguation method using the fine-grained word sense distinctions of the English SENSEVAL-2 data set [6].

To our knowledge, this is the first time a word sense disambiguation method using an independent large-scale representation of semantic space is reported.

The rest of this article is organized as follows: First we briefly discuss the WEBSOM method, then introduce our disambiguation method in Section II. The SENSEVAL-2 corpus and its preprocessing are described in Section III and the disambiguation experiments and results in Section IV. Sections V and VI present the discussion and conclusion, respectively.

A. WEBSOM document maps

The WEBSOM method [1], [13] uses the Self-Organizing Map algorithm [14] to organize a large document collection onto a two-dimensional display called the map. Documents are encoded using the bag-of-words model with word weighting. For computational reasons the dimensionality of the representation is reduced using random projection. The cosine measure (dot product for normalized vectors) is used for measuring similarities between documents.

Documents similar in content are located near each other on the ordered map display. The visualized display with an HTML interface can be used for exploring the document collection. The SOM consists of a set of map units ordered on a two-dimensional lattice. By virtue of a model vector stored with each map unit, searches can be performed on the map by locating the most similar model vector for a new document or short context.

It has been shown that the WEBSOM method is applicable to various kinds of text collections, including very colloquial ones. The method is scalable, as is shown by the organization of the map of nearly 7 million patent abstracts [1]. For our disambiguation experiments we use this existing large document map with 1,002,240 map units.

When preprocessing the text, mathematical symbols and numbers were converted to dummy symbols. Originally there were 733,179 different tokens in the document collection, but a cutoff threshold was applied leaving out words occurring less than 50 times, as well as 1,355 frequent words on a stop list. The words were reduced to their base forms and the remaining number of types was 43,222 words. Each document was positioned in a 43,222-dimensional space. For efficiency of map creation, the dimensionality of the vector space was condensed to 500 by the random projection procedure, which has been shown to retain the information of the original high-dimensional space while introducing only a small amount of random noise [15]. [1]

II. Word sense disambiguation method

Let us assume that there is a vast general purpose document collection containing a representative sample of many different areas of human endeavor, discussions or other texts representing varying discourses and different topics. One can then organize such a collection on a large document map and thus obtain a representation of semantic space.

Such a representation of semantic space can be used in word sense disambiguation as follows: A word in context is treated as a small document. As a *calibration* step, each sample, i.e., a sense-tagged word in context, is encoded as a document vector and positioned on the best-matching map location corresponding to that word and its context. The word and its context are both used

in the matching.

When an ambiguous word needs disambiguating the word in context is processed identically as the tagged samples, and the best-matching location in semantic space is found. The tag for the word is selected by a majority vote among the k nearest sense-tagged samples of the same word on the map lattice, cf. Figure 1. In case of a tie between several senses, the globally most frequent of them is chosen. A restriction[1] on the maximum allowed distance d_{max} is set; if a sample lies outside of this limit, it is not considered in the vote. If no samples lie near enough (within d_{max}), instead of the local decision strategy a global strategy is applied: a majority vote is taken among all the sense tagged samples of that word.

The distance between two words in context is defined as the map lattice distance between their locations on the map. The distance between neighboring map units is one[2].

A variant of the method uses fuzzy positions on the WEBSOM map. A fuzzy position consists of the l best matching map locations corresponding to the word in context. For small l (e.g. $l < 10$) these locations are generally near each other, but especially for larger l several clusters may emerge. In this variant, the distance between words w_1 and w_2 is defined as the minimum distance between the map locations corresponding to these two words.

Once a suitable map exists, the computational complexity of the method is low and is dominated by the search for the best matching location for the calibration and test samples. The search requires $\mathcal{O}(m * w * M)$ operations, where m is the number of calibration and test samples, w is the average number of words in each sample, and M is the number of map units of the map. When the best matching units have been located on the map, the disambiguation requires an additional $\mathcal{O}(n*k)$ operations, where n is the number of test samples, and k is the average number of nearest neighbors to be processed for each sample. For each test sample a constant time is needed for locating the nearest neighboring sense-tagged samples, if an index is used. The fuzzy variant of the disambiguation requires $\mathcal{O}(n*k*l^2)$ operations, where n and k are as before and l is the number of map locations for each sample.

III. Data and preprocessing

A. Data set

In order to compare the method with other systems we used the data set from the English lexical sample task of the SENSEVAL-2 competition [6]. The data set consisted of 8611 training samples and 4328 test samples of 73 words in context.

[1] The restriction on the maximum distance to the nearest neighbors is motivated by the fact that the WEBSOM map is likely to be only a locally accurate projection of the document clusters in the multi-dimensional semantic space.

[2] Strictly speaking this is true only for the rectangular SOM lattice; for the hexagonal lattice the calculation is slightly more complicated. In any case, the same distance calculation has been used as in defining the map neighborhood function in SOM learning, cf. [14]

The lexicon used for the sense inventory was WordNet 1.7 and the word instances were mostly from the British National Corpus with some from the Wall Street Journal. Inter-tagger-agreement was 85.5% [16]. There were between 70 and 455 instances per word[3] (divided 2:1 between training data and test data) with the averages of 121:61 for nouns, 91:46 for adjectives and 122:62 for verbs.

The words in context comprised 29 nouns (art, authority, bar, bum, chair, channel, child, church, circuit, day, detention, dyke, facility, fatigue, feeling, grip, hearth, holiday, lady, material, mouth, nation, nature, post, restraint, sense, spade, stress, yew), 15 adjectives (blind, colourless, cool, faithful, fine, fit, free, graceful, green, local, natural, oblique, simple, solemn, vital) and 29 verbs (begin, call, carry, collaborate, develop, draw, dress, drift, drive, face, ferret, find, keep, leave, live, match, play, pull, replace, see, serve, strike, train, treat, turn, use, wander, wash, work).

Examples of two noun senses for the word church are the building sense: "The little church has suffered from the rigours of time and town planning, normally a lethal combination, and survived.", and the congregation sense: "On the other hand to disassociate the Church from the Kingdom breaks the nerve-cord of hope and destroys the community of commitment to Christ as Saviour and Lord."

The words in the training samples had an average of 10.9 senses (8.7 senses in the test samples). Nouns had 8.2 (6.7) senses, adjectives 4.4 (4.1) and verbs 15.6 (12.2) on the average. The high number of senses is partly due to the fact that multi-word expressions are counted as separate senses of single words. There were 105 (67) compound nouns and 155 (106) phrasal verbs, which accounted for 8.5 % (7.5 %) of the noun and 9.7 % (10.1 %) of the verb data, respectively. Typically a compound noun would have only one sense and a phrasal verb would have one or two senses. Without the compounds and phrasal verbs the noun and verb sense ambiguity was 4.6 (4.4) and 10.2 (8.5) senses per word, respectively.

B. Preprocessing of the sense inventory

Spotting multi-word expressions was left to preprocessing with a phrase filter. For example, the word church was 4-ways ambiguous, but the collocation 'church bell' had only one sense. Without phrase filtering the word church would in this case have been regarded as 5-ways ambiguous. Due to the phrase filter a compound word or phrasal verb would sometimes be falsely recognized, e.g. "The first ladies' mile in ...", where the correct interpretation is the 'first mile' and not the 'first lady'. The falsely recognized multi-word expressions were so few (< 1 %) in the SENSEVAL training data that this was not considered important for the experiment.

After the phrase filter the words had an average of 7.2 senses per word in context in the training data (7.4

[3] The aim of the organizers was to have $75+15n+6m$ instances for each word, where n is the number of senses and m is the number of multi-word expressions for a word.

senses per word in the testing data). Nouns had 5.5 (5.6) senses, adjectives 6.1 (6.1) and verbs 9.6 (9.8) on the average.

C. Baselines

The most frequent sense baseline, which would be achieved by always selecting the most frequent of the candidate senses of a word in context, was correct in 47.6 % of the test samples. The expectation value baseline, which would be obtained by equally distributed random selection of a word sense from a full sense inventory of a word, provided 14 % correct sense assignments in the test data.

The phrase filter changes the baselines, because the full range of senses need no longer be considered. In the preprocessing we left out the proper noun and unknown senses of the words. We also discarded samples with more than one tag in the training data. This already changed the expectation value and most frequent sense baseline to 28 % and 51.3 %, respectively. In addition, the phrase filter assigned the compound word and phrasal verb senses only to words identified as compounds and phrasal verbs. This further increased the most frequent sense baseline to 52.3 %.

We realized that our phrase filter introduces approximately 4.1 % errors in the test data by not correctly identifying all compounds or phrasal verbs. This is quite a high percentage, but no time was spent on fixing the ad hoc phrase filter. The 181 faulty samples were disregarded. The most frequent sense baseline for the remaining 4147 samples was 55.1 %.

IV. Disambiguation experiments

We used the training samples of the SENSEVAL-2 data, which were disambiguated words in context, for calibrating the WEBSOM map. For the parameter selection we used 10-fold cross-validation on the calibration data to find the best-performing parameter combinations, which were then used for disambiguating the test data.

A. Parameter selection

A.1 Extent of the disambiguation context

The task was to assign the proper word sense to an ambiguous word in context. As the context of a word we used alternatively a paragraph, a sentence, a clause, or a 7-word window in which the word occurred. The best performance in the initial study was achieved when the sentence context was used both for calibration and disambiguation.

A.2 Selection of k

The ambiguous word in context was assigned the same sense as the most frequent sense of the k nearest sense-disambiguated neighbors. The parameter k was varied between 1 ... 20 and d_{max} between 1 ... 10. Maximum performance was achieved when $k = 3$ and $d_{max} = 2$.

A.3 Fuzzy vs. crisp positions on the map

We also evaluated the method using fuzzy positions on the WEBSOM map. A fuzzy position consisted of l best locations chosen in order of how well they matched the word in context. We evaluated the method using 1, 2, 5, 10 or 15 best locations for each word in context. Fuzzy locations did not improve the overall result, so we chose $l = 1$.

B. Test results

The final test results, were obtained using a separate test data set, namely the English lexical task test corpus of the SENSEVAL-2 competition. The method was evaluated on the fine-grained sense distinctions in the SENSEVAL test corpus yielding an overall performance of (2341/4147) 56.45 % correct results with a standard deviation of 0.77 %. This was 65.3±1.8 % for adjectives, 59.6±1.2 % for nouns and 46.9±1.2 % for verbs.

If we include the 181 errors committed by the phrase filter, we obtain an overall result of (2341/4328) 54.09±0.76 %.

C. Importance of the result

The results are statistically significant compared to their most frequent sense baselines on the 5 % significance level. A few methods in the SENSEVAL-2 competition attempted only a small fraction of the test samples and achieved very good results on those fractions. The supervised methods attempting >97 % of the test samples achieved an overall performance between 64.2-42.1 %. The fully unsupervised methods attempting >98 % achieved 40.2-22.0 %. The proposed method would have attained position 14 among all the 28 participants of the SENSEVAL-2 competition.

The result of the proposed method is remarkable, considering that the semantic ordering of the map is formed in a fully unsupervised manner, and optimized for text exploration, not for word sense disambiguation. It can be expected that if the application task is taken into account in the construction of the map, even better disambiguation performance may be obtained. In this sense, the current work is only a preliminary study of the potential of the method for the word sense disambiguation problem. Many of the participants in the SENSEVAL-2 competition were complex hybrid systems specifically optimized for that particular application. Our results show that a suitable representation of semantic space can be created independently of the domain and the application.

V. Discussion

As mentioned in Section II, the map was used only for finding neighbors within a d_{max} distance of a test word in context. If no neighbors were within reach, the word in context was disambiguated with a global most frequent sense vote. With the parameters determined by the cross-validation, the map was used for disambiguating 998 test samples. The portion of correct results among the samples, where the map was applied, was 61.9±1.54 % with a most frequent sense baseline of

56.6 %. It might seem problematic that the map was used only for 998 of the 4157 context data samples, but this is mainly due to the small amount of calibration data. In [17], a supervised method uses a nearest neighbor algorithm directly on the training samples with 100-1300 training samples per ambiguous word with the conclusion that results improve considerably as the amount of training data increases.

The map used in the evaluation was created from 7 million documents. Some of the words in the SENSEVAL-2 data were still not in the vocabulary when the WEBSOM map was created. Despite this e.g. the adjective *free*, which was not in the training vocabulary of the document map, performed statistically significantly better than its individual baseline in the word sense disambiguation. This seems to support the assumption that the map has induced a topical and discourse ordering of contexts, which is used as a lever in the word sense disambiguation task.

The document map was created from patent abstracts, which means that the domain and the genre of the texts were fairly different from the evaluation data. Even if the word was in the patent vocabulary and the word sense distinction was of a topical nature, some senses might not be represented at all on the map. Examples of this are the building and the congregation senses of the word church. There are no patents pertaining to the congregation sense of church, but there are many relating to the building sense. However, the result for *church/N*[4] was not statistically significantly lower than the baseline. Only the words *bum/N* and *begin/V* perform statistically significantly below the baseline, whereas the words *authority/N, channel/N, circuit/N, mouth/N, spade/N, restraint/N, fine/A, free/A, dress/V, drive/V, play/V, replace/V, serve/V* and *train/V* all perform statistically significantly above the baseline. We see that all of the studied parts of speech are represented, but words with senses related to the patent domain perform best in the disambiguation task, which is to be expected. It might be interesting as a future direction to study the effects of a smaller, more general purpose document collection organized as a document map.

The WEBSOM map aims at representing topic and discourse information and disregards morphosyntactic information. This seems to work for the disambiguation contexts of the words mentioned above. However, in some cases morphosyntactic information provides important cues for disambiguation. An even better method would perhaps be obtained by combining the strengths of the topically ordered WEBSOM map with some form of morphosyntactic analysis. A simple way of doing this would be to let a linguistic preprocessor select the senses to be disambiguated on the basis of the local context leaving only the topic and discourse related ambiguities for the word-sense disambiguation in semantic space.

As reported e.g. in [1] the WEBSOM map has two different means for locating a map position for a word in context: document search (or *content-addressable search*) and keyword search. In the future our intent is to examine whether the keyword search mode would provide an even better basis for the disambiguation than the currently used document search.

[4] */A* means adjective, */V* verb and */N* noun

VI. Conclusion

We have shown that an independent large-scale representation of semantic space can be utilized in word sense disambiguation. The current study also shows that the distributed knowledge of a WEBSOM map can be used as a representation of semantic space.

Based on the experiments it is apparent that the existing large WEBSOM map distilled from a massive patent abstract collection provided a successful representation of semantic space, even though the map and the original document collection were not initially intended for word sense disambiguation. Constructing a map of a large balanced document collection might yield further improvements.

What is even more interesting, the current work succeeds in showing experimentally what has earlier been only a hypothesis, namely that the topical ordering obtained using the WEBSOM method truly reflects a general property of semantic representations. Moreover, the word sense disambiguation problem as well as other NLP problems dealing with lexical semantics can provide us indirect information regarding the properties of the cognitive semantic apparatus that humans utilize. The fact that a property such as topical ordering manifests itself across problem types and application domains is thus remarkable from the point of view of cognitive modeling: it provides strong indirect evidence in support of the hypothesis that a similar ordering may be employed also by the human brain.

Acknowledgement

We are grateful to the WEBSOM team at the Neural Networks Research Centre of the Helsinki University of Technology for the permission to use the patent abstract map for this experiment.

References

[1] Teuvo Kohonen, Samuel Kaski, Krista Lagus, Jarkko Salojärvi, Vesa Paatero, and Antti Saarela, "Organization of a massive document collection," *IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery*, vol. 11, no. 3, pp. 574–585, May 2000.

[2] Nancy Ide and Jean Veronis, "Introduction to the special issue on word sense disambiguation: The state of the art," *Computational Linguistics*, vol. 24, no. 1, pp. 1–40, 1998.

[3] Christopher D. Manning and Hinrich Schütze, *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, Massachusetts, 1999.

[4] David Yarowsky, "Unsupervised word-sense disambiguation rivaling supervised methods," in *Proceedings of the 33rd Annual Meeting of the Association for Computational Lainguistics (ACL '95)*, Cambridge, MA, 1995, pp. 189–196.

[5] David Martinez and Eneko Agirre, "One sense per collocation and genre/topic variations," in *Proceedings of the Joint SIGDAT Conference on Empirical Methods in*

Natural Language Processing and Very Large Corpora, Hong Kong, 2000.

[6] SENSEVAL-2, "Training and testing corpora," *http://www.cis.upenn.edu/∼cotton/senseval/corpora.tgz*, 2001.

[7] Gerard Escudero, Lluís Màrquez, and German Rigau, "A comparison between supervised learning algorithms for word sense disambiguation," in *Proceedings of CoNLL-2000 and LLL-2000*, Claire Cardie, Walter Daelemans, Claire Nedellec, and Erik Tjong Kim Sang, Eds. 2000, pp. 31–36, Lisbon, Portugal.

[8] E. Agirre, G. Rigau, J. Atserias, and L. Padró, "Combining supervised and unsupervised lexical knowledge methods for word sense disambiguation," *Computers and Humanities*, vol. 34, no. 1-2, 2000.

[9] Hinrich Schütze, "Automatic word sense discrimination.," *Computational Linguistics*, vol. 24, no. 1, pp. 97–123, 1998.

[10] Will Lowe, "Towards a theory of semantic space," in *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society*, J. D. Moore and K. Stenning, Eds., Mahwah NJ, 2001, pp. 576–581, Lawrence Erlbaum Associates.

[11] M. Steyvers and J. B. Tenenbaum, "The large-scale structure of semantic networks: statistical analyses and a model of semantic growth," *Cognitive Science*, to appear.

[12] Thomas Martinetz and Klaus Schulten, "Topology representing networks," *Neural Networks*, vol. 7, no. 3, pp. 507–522, 1994.

[13] Timo Honkela, Samuel Kaski, Krista Lagus, and Teuvo Kohonen, "Newsgroup exploration with WEBSOM method and browsing interface," Tech. Rep. A32, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland, 1996.

[14] Teuvo Kohonen, *Self-Organizing Maps*, Springer, Berlin, 1995.

[15] Samuel Kaski, "Dimensionality reduction by random mapping: Fast similarity computation for clustering," in *Proceedings of IJCNN'98, International Joint Conference on Neural Networks*, vol. 1, pp. 413–418. IEEE Service Center, Piscataway, NJ, 1998.

[16] Adam Kilgariff, "English lexical sample task description," *http://www.itri.bton.ac.uk/events/senseval /englexsamp.ps*, 2001.

[17] Hwee Tou Ng, "Getting serious about word sense disambiguation," in *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, Washington, D.C., USA, 1997, pp. 1–7.