Getting Serious about Word Sense Disambiguation

Hwee Tou Ng
Defence Science Organisation
20 Science Park Drive
Singapore 118230
Republic of Singapore
nhweetou@dso.gov.sg

Abstract

Recent advances in large-scale, broad coverage part-of-speech tagging and syntactic parsing have been achieved in no small part due to the availability of large amounts of online, human-annotated corpora. In this paper, I argue that a large, human sense-tagged corpus is also critical as well as necessary to achieve broad coverage, high accuracy word sense disambiguation, where the sense distinction is at the level of a good desk-top dictionary such as WordNet. Using the sense-tagged corpus of 192,800 word occurrences reported in (Ng and Lee, 1996), I examine the effect of the number of training examples on the accuracy of an exemplar-based classifier versus the base-line, most-frequent-sense classifier. I also estimate the amount of human sense-tagged corpus and the manual annotation effort needed to build a large-scale, broad coverage word sense disambiguation program which can significantly outperform the most-frequent-sense classifier. Finally, I suggest that intelligent example selection techniques may significantly reduce the amount of sense-tagged corpus needed and offer this research problem as a fruitful area for word sense disambiguation research.

1 Introduction

Much recent research in the field of natural language processing (NLP) has focused on an empirical, corpus-based approach (Church and Mercer, 1993). The high accuracy achieved by a corpus-based approach to part-of-speech tagging and noun phrase parsing, as demonstrated by (Church, 1988), has inspired similar approaches to other problems in natural language processing, including syntactic parsing and word sense disambiguation (WSD).

The availability of large quantities of part-of-speech tagged and syntactically parsed sentences like the Penn Treebank corpus (Marcus, Santorini, and Marcinkiewicz, 1993) has contributed greatly to the development of robust, broad coverage part-of-speech taggers and syntactic parsers. The Penn Treebank corpus contains a sufficient number of part-of-speech tagged and syntactically parsed sentences to serve as adequate training material for building broad coverage part-of-speech taggers and parsers.

Unfortunately, an analogous sense-tagged corpus large enough to achieve broad coverage, high accuracy word sense disambiguation is not available at present. In this paper, I argue that, given the current state-of-the-art capability of automated machine learning algorithms, a supervised learning approach using a large sense-tagged corpus is a viable way to build a robust, wide coverage, and high accuracy WSD program. In this view, a large sense-tagged corpus is critical as well as necessary to achieve broad coverage, high accuracy WSD.

The rest of this paper is organized as follows. In Section 2, I briefly discuss the utility of WSD in practical NLP tasks like information retrieval and machine translation. I also address some objections to WSD research. In Section 3, I examine the size of the training corpus on the accuracy of WSD, using a corpus of 192,800 occurrences of 191 words hand tagged with WordNet senses (Ng and Lee, 1996). In Section 4, I estimate the amount of human sense-tagged corpus and the manual annotation effort needed to build a broad coverage, high accuracy WSD program. Finally, in Section 5, I suggest that intelligent example selection techniques may significantly reduce the amount of sense-tagged corpus needed and offer this research problem as a fruitful area for WSD research.

2 The Utility of Word Sense Disambiguation

Although there is agreement in general about the utility of WSD within the NLP community, I will briefly address some objections to WSD in this section. To justify the investment of manpower and time to gather a large sense-tagged corpus, it is important to examine the benefits brought about by WSD.

Information retrieval (IR) is a practical NLP task where WSD has brought about improvement in accuracy. When tested on some standard IR test collection, the use of WSD improves precision by about 4.3% (from 29.9% to 34.2%) (Schütze and Pedersen, 1995). The work of (Dagan and Itai, 1994) has also successfully used WSD to improve the accuracy of machine translation. These examples clearly demonstrate the utility of WSD in practical NLP applications.

In this paper, by word sense disambiguation, I mean identifying the correct sense of a word in context such that the sense distinction is at the level of a good desk-top dictionary like WordNet (Miller, 1990). I only focus on content word disambiguation (*i.e.*, words in the part of speech noun[1], verb, adjective and adverb). This is also the task addressed by other WSD research such as (Bruce and Wiebe, 1994; Miller et al., 1994). When the task is to resolve word senses to the fine-grain distinction of WordNet senses, the accuracy figures achieved are generally not very high (Miller et al., 1994; Ng and Lee, 1996). This indicates that WSD is a challenging task and much improvement is still needed.

However, if one were to resolve word sense to the level of *homograph*, or coarse sense distinction, then quite high accuracy can be achieved (in excess of 90%), as reported in (Wilks and Stevenson, 1996). Similarly, if the task is to distinguish between binary, coarse sense distinction, then current WSD techniques can achieve very high accuracy (in excess of 96% when tested on a dozen words in (Yarowsky, 1995)). This is to be expected, since homograph contexts are quite distinct and hence it is a much simpler task to disambiguate among a small number of coarse sense classes. This is in contrast to disambiguating word senses to the refined senses of WordNet, where for instance, the average number of senses per noun is 7.8 and the average number of senses per verb is 12.0 for the set of 191 most ambiguous words investigated in (Ng and Lee, 1996).

We can readily collapse the refined senses of WordNet into a smaller set if only a coarse (homographic) sense distinction is needed, say for some NLP applications. Indeed, the WordNet software has an option for grouping noun senses into a smaller number of sense classes. WSD techniques that work well for refined sense distinction will apply equally to homograph disambiguation. That is, if we succeed in working on the harder WSD task of resolution into refined senses, the same techniques will also work on the simpler task of homograph disambiguation.

A related objection to WSD research is that the sense distinction made by a good desk-top dictionary like WordNet is simply too refined, to the point that two humans cannot genuinely agree on the most appropriate sense to assign to some word occurrence (Kilgarriff, 1996). This objection has some merits. However, the remedy is not to throw out word senses completely, but rather to work on a level of sense distinction that is somewhere in between homograph distinction and the refined WordNet sense distinction. The existing lumping of noun senses in WordNet into coarser sense groups is perhaps a good compromise.

However, in the absence of well accepted guidelines for making an appropriate level of sense distinction, using the sense classification given in WordNet, an on-line, publicly available dictionary, seems a natural choice. Hence, I believe that using the current WordNet sense distinction to build a sense-tagged corpus is a reasonable approach to go forward. In any case, if some aggregation of senses into coarser grouping is done in future, this can be readily incorporated into my proposed sense-tagged corpus which uses the refined sense distinction of WordNet.

In the rest of this paper, I will assume that broad coverage, high accuracy WSD is indeed useful in practical NLP tasks, and that resolving senses to the refined level of WordNet is a worthwhile task to pursue.

3 The Effect of Training Corpus Size

A number of past research work on WSD, such as (Leacock et al., 1993; Bruce and Wiebe, 1994; Mooney, 1996), were tested on a small number of words like "line" and "interest". Similarly, (Yarowsky, 1995) tested his WSD algorithm on a dozen words. The sense-tagged corpus Semcor, prepared by (Miller et al., 1994), contains a substantial subset of the Brown corpus tagged with the refined senses of WordNet. However, as reported in (Miller et al., 1994), there are not enough training examples per word in Semcor to yield a broad coverage, high accuracy WSD program, due to the fact that sense tagging is done on every word in a

[1] I will only focus on common noun in this paper and ignore proper noun.

running text in SEMCOR.

To overcome this data sparseness problem of WSD, I initiated a mini-project in sense tagging and collected a corpus in which 192,800 occurrences of 191 words have been manually tagged with senses of WORDNET (Ng and Lee, 1996). These 192,800 word occurrences consist of only 121 nouns and 70 verbs which are the most frequently occurring and most ambiguous words of English.[2]

To investigate the effect of the number of training examples on WSD accuracy, I ran the exemplar-based WSD algorithm LEXAS on varying number of training examples to obtain learning curves for the 191 words (details of LEXAS are described in (Ng and Lee, 1996)). For each word, 10 random trials were conducted and the accuracy figures were averaged over the 10 trials. In each trial, 100 examples were randomly selected to form the test set, while the remaining examples (randomly shuffled) were used for training. LEXAS was given training examples in multiples of 100, starting with 100, 200, 300, ... training examples, up to the maximum number of training examples (in a multiple of 100) available in the corpus.

Note that each word w (of the 191 words) can have a different number of sense-tagged occurrences in our corpus. From the combination of Brown corpus (1 million words) and Wall Street Journal corpus (2.5 million words), up to 1,500 sentences each containing an occurrence of the word w are extracted from the combined corpus, with each sentence containing a sense-tagged occurrence of w. When the combined corpus has less than 1,500 occurrences of w, the maximum number of available occurrences of w is used. For instance, while 137 words have at least 600 occurrences in the combined corpus, only a subset of 43 words has at least 1400 occurrences. Figure 1 and 2 show the learning curves averaged over these 43 words and 137 words with at least 1300 and 500 training examples, respectively. Each figure shows the accuracy of LEXAS versus the base-line, most-frequent-sense classifier.

Both figures indicate that WSD accuracy continues to climb as the number of training examples increases. They confirm that all the training examples collected in our corpus are effectively utilized by LEXAS to improve its WSD performance. In fact, it appears that for this set of most ambiguous words of English, more training data may be beneficial to further improve WSD performance.

I also report here the evaluation of LEXAS on two subsets of test sentences of our sense-tagged corpus, as shown in Table 1.

Test set	Sense 1	Most Frequent	LEXAS
BC50	40.5%	47.1%	58.7%
WSJ6	44.8%	63.7%	75.2%

Table 1: Evaluation of LEXAS

The two test sets, BC50 and WSJ6, are the same as those reported in (Ng and Lee, 1996). BC50 consists of 7,119 occurrences of the 191 words that occur in 50 text files of the Brown corpus. The second test set, WSJ6, consists of 14,139 occurrences of these 191 words that occur in 6 text files of the Wall Street Journal corpus.

The performance figures of LEXAS in Table 1 are higher than those reported in (Ng and Lee, 1996). The classification accuracy of the nearest neighbor algorithm used by LEXAS (Cost and Salzberg, 1993) is quite sensitive to the number of nearest neighbors used to select the best matching example. By using 10-fold cross validation (Kohavi and John, 1995) to automatically pick the best number of nearest neighbors to use, the performance of LEXAS has improved.

4 Word Sense Disambiguation in the Large

In (Gale et al., 1992), it was argued that any wide coverage WSD program must be able to perform significantly better than the most-frequent-sense classifier to be worthy of serious consideration. The performance of LEXAS as indicated in Table 1 is significantly better than the most-frequent-sense classifier for the set of 191 words collected in our corpus. Figure 1 and 2 also confirm that all the training examples collected in our corpus are effectively utilized by LEXAS to improve its WSD performance. This is encouraging as it demonstrates the feasibility of building a wide coverage WSD program using a supervised learning approach.

Unfortunately, our corpus only contains tagged senses for 191 words, and this set of words does not constitute a sufficiently large fraction of all occurrences of content words in an arbitrarily chosen unrestricted text. As such, our sense-tagged corpus is still not large enough to enable the building of a wide coverage, high accuracy WSD program that can significantly outperform the most-frequent-sense classifier over all content words encountered in an arbitrarily chosen unrestricted text.

This brings us to the question: how much data do we need to achieve wide coverage, high accuracy WSD?

[2]This corpus is scheduled for release by the Linguistic Data Consortium (LDC). Contact the LDC at ldc@unagi.cis.upenn.edu for details.

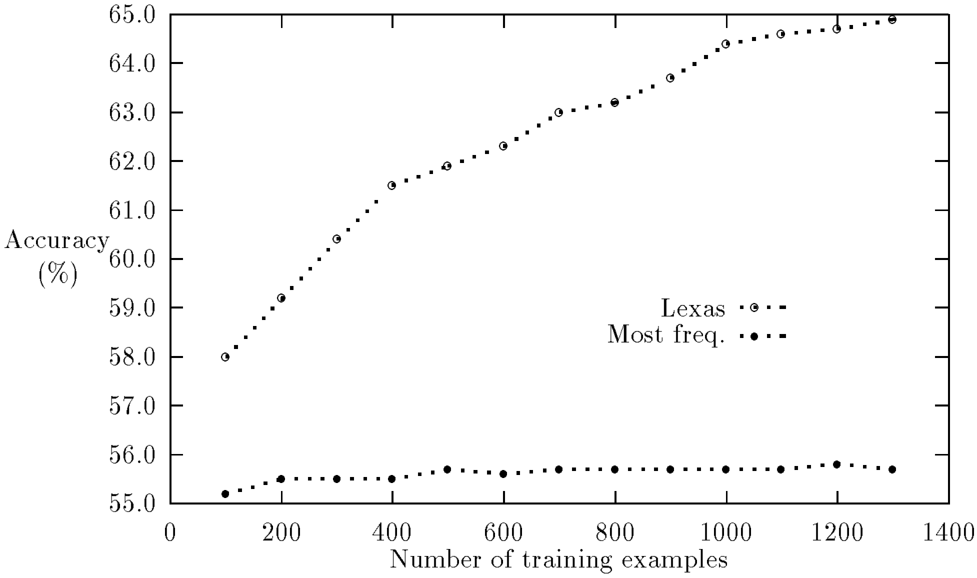

Figure 1: Effect of number of training examples on WSD accuracy averaged over 43 words with at least 1300 training examples

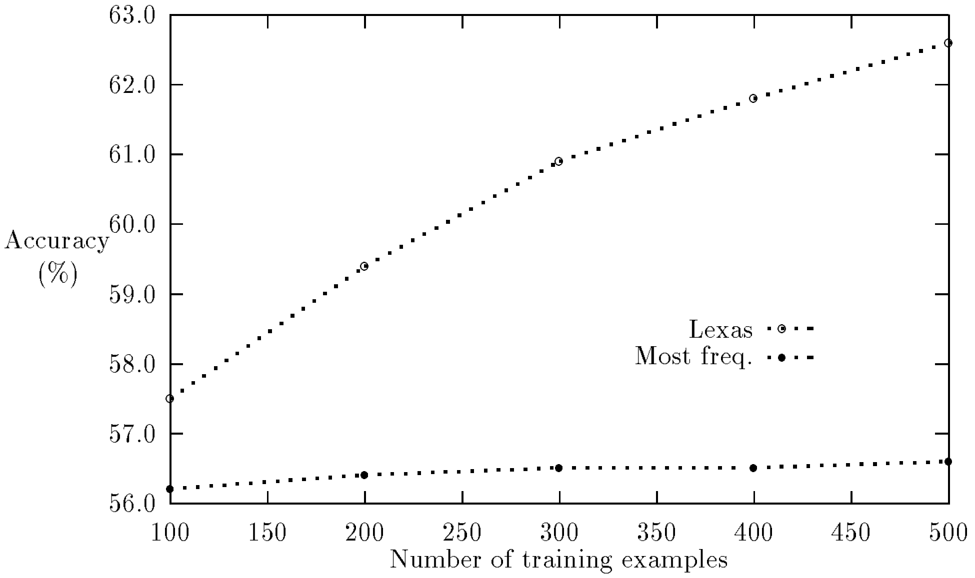

Figure 2: Effect of number of training examples on WSD accuracy averaged over 137 words with at least 500 training examples

POS	80%	90%	95%	99%
noun	975	1776	2638	4510
verb	242	550	926	1806
adj	374	769	1286	2384
adv	36	76	128	269
sum	1627	3171	4978	8969

Table 2: Number of polysemous words in each part of speech making up the top 80%, ..., 99% of word occurrences in the Brown corpus.

POS	80%	90%	95%	99%
noun	472	946	1520	3130
verb	203	429	707	1487
adj	171	402	761	1748
adv	35	69	104	206
sum	881	1846	3092	6571

Table 3: Number of polysemous words in each part of speech making up the top 80%, ..., 99% of word occurrences in the Wall Street Journal corpus.

To shed light on this question, it is instructive to examine the distribution of words and their occurrence frequency in a large corpus. Table 2 lists the number of polysemous words in each part of speech making up the top 80%, ..., top 99% of word occurrences in the Brown corpus, where the polysemous words are ordered in terms of their occurrence frequency from the most frequently occurring word to the least frequently occurring word. For example, Table 2 indicates that when the polysemous nouns are ordered from the most frequently occurring noun to the least frequently occurring noun, the top 975 polysemous nouns constitute 80% of all noun occurrences in the Brown corpus. This 80% of all noun occurrences include all nouns in the Brown corpus that are monosemous (about 15.4%) and all rare nouns in the Brown corpus that do not appear in WordNet and hence have no valid sense definition (about 3.3%) (*i.e.*, the remaining 20% noun occurrences are all polysemous). Table 3 lists the analogous statistics for the Wall Street Journal corpus.

It is also the case that the last 5%–10% of polysemous words in a corpus have only a small number of distinct senses on average. Table 4 lists the average number of senses per polysemous word in the Brown corpus for the top 80%, ..., top 99%, and the bottom 20%, ..., bottom 1% of word occurrences, where the words are again ordered from the most frequently occurring word to the least frequently occurring word. For example, the average number of senses per polysemous noun is 5.14 for the nouns which account for the top 80% noun occurrences in the Brown corpus. Similarly, the average number of senses per polysemous noun is 2.86 for the polysemous nouns which account for the bottom 20% of noun occurrences in the Brown corpus. Table 5 lists the analogous statistics for the Wall Street Journal corpus.

Table 2 and 3 indicate that a sense-tagged corpus collected for 3,200 words will cover at least 90% of all (content) word occurrences in the Brown corpus, and at least 95% of all (content) word occurrences in the Wall Street Journal corpus. From Table 4, the average number of senses per polysemous word in the Brown corpus for the remaining 10% word occurrences is only 3.15 or less. Similarly, from Table 5, the average number of senses per polysemous word in the Wall Street Journal corpus for the remaining 5% word occurrences is only 3.10 or less. For these remaining polysemous words which account for the last 5%–10% word occurrences with an average of about 3 senses per word, we can always assign the most frequent sense as a first approximation in building our wide coverage WSD program.

Based on these figures, I estimate that a sense-tagged corpus of 3,200 words is sufficient to build a broad coverage, high accuracy WSD program capable of significantly outperforming the most-frequent-sense classifier on average over all content words appearing in an arbitrary, unrestricted English text. Assuming an average of 1,000 sense-tagged occurrences per word, this will mean a corpus of 3.2 million sense-tagged word occurrences. Assuming human sense tagging throughput at 200 words, or 200,000 word occurrences, per man-year (which is the approximate human tagging throughput of my completed sense-tagging mini-project), such a corpus will require about 16 man-years to construct.

Given the benefits of a wide coverage, high accuracy and domain-independent WSD program, I believe it is justifiable to spend the 16 man-years of human annotation effort needed to construct such a sense-tagged corpus.

5 Can We Do Better?

My estimate of the amount of human annotation effort needed can be considered as an upper bound on the manual effort needed to construct the necessary sense-tagged corpus to achieve wide coverage WSD. It may turn out that we can achieve our goal with much less annotation effort.

Recent work on intelligent example selection techniques suggest that the quality of the examples used for supervised learning can have a large impact on the classification accuracy of the induced classifier. For example, in (Engelson and Dagan, 1996),

POS	top 80%	top 90%	top 95%	top 99%	bottom 20%	bottom 10%	bottom 5%	bottom 1%
noun	5.14	4.48	4.07	3.51	2.86	2.71	2.59	2.44
verb	8.75	6.89	5.77	4.53	3.43	3.15	2.94	2.67
adj	5.87	4.75	4.08	3.47	2.86	2.72	2.63	2.44
adv	4.22	3.79	3.48	2.96	2.55	2.46	2.38	2.31

Table 4: Average number of senses per polysemous word in the Brown corpus for the top 80%, ..., top 99%, and the bottom 20%, ..., bottom 1% of word occurrences.

POS	top 80%	top 90%	top 95%	top 99%	bottom 20%	bottom 10%	bottom 5%	bottom 1%
noun	5.44	4.89	4.50	3.83	3.08	2.95	2.83	2.60
verb	8.72	7.13	6.19	4.75	3.52	3.30	3.10	2.87
adj	6.13	5.33	4.63	3.76	3.09	2.95	2.81	2.60
adv	4.00	3.67	3.55	3.14	2.62	2.56	2.48	2.37

Table 5: Average number of senses per polysemous word in the Wall Street Journal corpus for the top 80%, ..., top 99%, and the bottom 20%, ..., bottom 1% of word occurrences.

committee-based sample selection is applied to part-of-speech tagging to select for annotation only those examples that are the most informative, and this avoids redundantly annotating examples. Similarly, in (Lewis and Catlett, 1994), uncertainty sampling of training examples achieved better accuracy than random sampling of training examples for a text categorization application.

Intelligent example selection for supervised learning is an important issue of machine learning in its own right. I believe it is of particular importance to investigate this issue in the context of word sense disambiguation, as the payoff is high, given that a large sense tagged corpus is currently not available and remains one of the most critical bottlenecks in achieving wide coverage, high accuracy WSD.

References

Rebecca Bruce and Janyce Wiebe. 1994. Word-sense disambiguation using decomposable models. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, New Mexico.

Kenneth Church. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing (ANLP)*, pages 136–143.

Kenneth W. Church and Robert L. Mercer. 1993. Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, 19(1):1–24.

Scott Cost and Steven Salzberg. 1993. A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning*, 10(1):57–78.

Ido Dagan and Alon Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596.

Sean P. Engelson and Ido Dagan. 1996. Minimizing manual annotation cost in supervised training from corpora. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 319–326.

William Gale, Kenneth Ward Church, and David Yarowsky. 1992. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, Newark, Delaware.

Adam Kilgarriff. 1996. "I don't believe in word senses". manuscript.

Ron Kohavi and George H. John. 1995. Automatic parameter selection by minimizing estimated error. In *Machine Learning: Proceedings of the Twelfth International Conference.*

Claudia Leacock, Geoffrey Towell, and Ellen Voorhees. 1993. Corpus-based statistical sense resolution. In *Proceedings of the ARPA Human Language Technology Workshop.*

David D. Lewis and Jason Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. In *Machine Learning: Proceedings of the Eleventh International Conference.*

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

George A. Miller, Ed. 1990. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–312.

George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. Using a semantic concordance for sense identification. In *Proceedings of the ARPA Human Language Technology Workshop*.

Raymond J. Mooney. 1996. Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. In *Proceedings of the First Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Hwee Tou Ng and Hian Beng 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)*, pages 40–47.

Hinrich Schütze and Jan O. Pedersen. 1995. Information retrieval based on word senses. In *Symposium on Document Analysis and Information Retrieval*.

Yorick Wilks and Mark Stevenson. 1996. The grammar of sense: Is word-sense tagging much more than part-of-speech tagging? In Computational Linguistics Eprint Archive, cmp-lg/9607028.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, Cambridge, Massachusetts.