Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL '98), 1998, Montreal, Canada, pp 1398-1402.

Word Sense Disambiguation using Optimised Combinations of Knowledge Sources

Yorick Wilks and **Mark Stevenson**
Department of Computer Science,
University of Sheffield,
Regent Court, 211 Portobello Street,
Sheffield, S1 4DP
United Kingdom
{yorick, marks}@dcs.shef.ac.uk

Abstract

Word sense disambiguation algorithms, with few exceptions, have made use of only one lexical knowledge source. We describe a system which performs word sense disambiguation on all content words in free text by combining different knowledge sources: semantic preferences, dictionary definitions and subject/domain codes along with part-of-speech tags, optimised by means of a learning algorithm. We also describe the creation of a new sense tagged corpus by combining existing resources. Tested accuracy of our approach on this corpus exceeds 92%, demonstrating the viability of all-word disambiguation rather than restricting oneself to a small sample.

1 Introduction

This paper describes a system that integrates a number of partial sources of information to perform word sense disambiguation (WSD) of content words in general text at a high level of accuracy.

The methodology and evaluation of WSD are somewhat different from those of other NLP modules, and one can distinguish three aspects of this difference, all of which come down to evaluation problems, as does so much in NLP these days. First, researchers are divided between a general method (that attempts to apply WSD to all the content words of texts, the option taken in this paper) and one that is applied only to a small trial selection of texts words (for example (Schütze, 1992) (Yarowsky, 1995)). These researchers have obtained very high levels of success, in excess of 95%, close to the figures for other "solved" NLP modules, the issue being whether these small word sample methods and techniques will transfer to general WSD over all content words.

Others, (eg. (Mahesh et al., 1997) (Harley and Glennon, 1997)) have pursued the general option on the grounds that it is the real task and should be tackled directly, but with rather lower success rates. The division between the approaches probably comes down to no more than the availability of gold standard text in sufficient quantities, which is more costly to obtain for WSD than other tasks. In this paper we describe a method we have used for obtaining more test material by transforming one resource into another, an advance we believe is unique and helpful in this impasse.

However, there have also been deeper problems about evaluation, which has led sceptics like (Kilgarriff, 1993) to question the whole WSD enterprise, for example that it is harder for subjects to assign one and only one sense to a word in context (and hence the produce the test material itself) than to perform other NLP related tasks. One of the present authors has discussed Kilgarriff's figures elsewhere (Wilks, 1997) and argued that they are not, in fact, as gloomy as he suggests. Again, this is probably an area where there is an "expertise effect": some subjects can almost certainly make finer, more intersubjective, sense distinctions than others in a reliable way, just as lexicographers do.

But there is another, quite different, source of unease about the evaluation base: everyone agrees that new senses appear in corpora that cannot be assigned to any existing dictionary sense, and this is an issue of novelty, not just one of the difficulty of discrimination. If that is the case, it tends to undermine the standard mark-up-model-and-test methodology of most recent NLP, since it will not then be possible to mark up sense assignment in advance against a dictionary if new senses are present. We shall not tackle this difficult issue further here, but press on towards experiment.

2 Knowledge Sources and Word Sense Disambiguation

One further issue must be mentioned, because it is unique to WSD as a task and is at the core of our approach. Unlike other well-known NLP modules, WSD seems to be implementable by a number of apparently different information sources. All the following have been implemented as the basis of experimental WSD at various times: part-of-speech, semantic preferences, collocating items or classes, thesaural or subject areas, dictionary definitions, synonym lists, among others (such as bilingual equivalents in parallel texts). These phenomena seem different, so how can they all be, separately or in combination, informational clues to a single phenomenon, WSD? This is a situation quite unlike syntactic parsing or part-of-speech tagging: in the latter case, for example, one can write a Cherry-style rule tagger or an HMM learning model, but there is no reason the believe these represent different types of information, just different ways of conceptualising and coding it. That seems not to be the case, at first sight, with the many forms of information for WSD. It is odd that this has not been much discussed in the field.

In this work, we shall adopt the methodology first explicitly noted in connection with WSD by (McRoy, 1992), and more recently (Ng and Lee, 1996), namely that of bringing together a number of partial sources of information about a phenomenon and combining them in a principled manner. This is in the AI tradition of combining "weak" methods for strong results (usually ascribed to Newell (Newell, 1973)) and used in the CRL-NMSU lexical work on the Eighties (Wilks et al., 1990). We shall, in this paper, offer a system that combines the three types of information listed above (plus part-of-speech filtering) and, more importantly, applies a learning algorithm to determine the optimal combination of such modules for a given word distribution; it being obvious, for example, that thesaural methods work for nouns better than for verbs, and so on.

3 The Sense Tagger

We describe a system which is designed to assign sense tags from a lexicon to general text. We use the *Longman Dictionary of Contemporary English* (LODCE)(Procter, 1978), which contains two levels of sense distinction: the broad homograph level and the more fine-grained level of sense distinction.

Our tagger makes use of several modules which perform disambiguation and these are of two types: *filters* and *partial taggers*. A filter removes senses from consideration, thereby reducing the complexity of the disambiguation task. Each partial tagger makes use of a different knowledge source from the lexicon and uses it to suggest a set of possible senses for each ambiguous word in context. None of these modules performs the disambiguation alone but they are combined to make use of all of their results.

3.1 Preprocessing

Before the filters or partial taggers are applied the text is tokenised, lemmatised, split into sentences and part-of-speech tagged using the Brill part-of-speech tagger (Brill, 1992).

Our system disambiguates only the content words in the text[1] (the part-of-speech tags assigned by Brill's tagger are used to decide which are content words).

[1]We define content words as nouns, verbs, adjectives and adverbs, prepositions are not included in this class.

3.2 Part-of-speech

Previous work (Wilks and Stevenson, 1998) showed that part-of-speech tags can play an important role in the disambiguation of word senses. A small experimentwas carried out on a 1700 word corpus taken from the *Wall Street Journal* and, using only part-of-speech tags, an attempt was made to find the correct LDOCE homograph for each of the content words in the corpus. The text was part-of-speech tagged using Brill's tagger and homographs whose part-of-speech category did not agree with the tags assigned by Brill's system were removed from consideration. The most frequently occuring of the remaining homographs was chosen as the sense of each word. We found that 92% of content words were assigned the correct homograph compared with manual disambiguation of the same texts.

While this method will not help us disambiguate within the homograph, since all senses which combine to form an LDOCE homograph have the same part-of-speech, it will help us to identify the senses completely innapropriate for a given context (when the homograph's part-of-speech disagrees with that assigned by a tagger).

It could be reasonably argued that this is a dangerous strategy since, if the part-of-speech tagger made an error, the correct sense could be removed from consideration. As a precaution against this we have designed our system so that if none of the dictionary senses for a given word agree with the part-of-speech tag then they are all kept (none removed from consideration).

There is also good evidence from our earlier WSD system (Wilks and Stevenson, 1997) that this approach works well despite the part-of-speech tagging errors, that system's results improved by 14% using this strategy. achieved 88% correct disambiguation to the LDOCE homograph using this strategy but only 74% without it.

3.3 Dictionary Definitions

(Cowie et al., 1992) used simulated annealing to optimise the choice of senses for a text, based upon their textual definition in a dictionary. The optimisation was over a simple count of words in common in definitions, however, this meant that longer definitions were preferred over short ones, since they have more words which can contribute to the overlap, and short definitions or definitions by synonym were correspondingly penalised. We attempted to solve this problem as follows. Instead of each word contributing one we normalise its contribution by the number of words in the definition it came from. The Cowie et. al. implementation returned one sense for each ambiguous word in the sentence, without any indication of the system's confidence in its choice, but, we have adapted the system to return a set of suggested senses for each ambiguous word in the sentence. We found that the new evaluation function led to an improvement in the algorithm's effectiveness.

3.4 Pragmatic Codes

Our next partial tagger makes use of the hierarchy of LDOCE pragmatic codes which indicate the likely subject area for a sense. Disambiguation is carried out using a modified version of the simulated annealing algorithm, and attempts to optimise the number of pragmatic codes of the same type in the sentence. Rather than processing over single sentences we optimise over entire paragraphs and only for the sense of nouns. We chose this strategy since there is good evidence (Gale et al., 1992) that nouns are best disambiguated by broad contextual considerations, while other parts of speech are resolved by more local factors.

3.5 Selectional Restrictions

LDOCE senses contain simple selectional restrictions for each content word in the dictionary. A set of 35 semantic classes are used, such as `H` = Human, `M` = Human male, `P` = Plant, `S` = Solid and so on. Each word sense for a noun is given one of these semantic types, senses for adjectives list the type which they expect for the noun they modify, senses for adverbs the type they expect of their modifier and verbs list between one and three types (depending on their transitivity) which are the expected semantic types of the verb's subject, direct object and indirect object. Grammatical links between verbs, adjectives and adverbs and the head noun of their arguments arer identified using a specially constructed shallow syntactic analyser (Stevenson, 1998).

The semantic classes in LDOCE are not provided with a hierarchy, but, Bruce and Guthrie (Bruce and Guthrie, 1992) manually identified hierarchical relations between the semantic classes, constructing them into a hierarchy which we use to resolve the restrictions. We resolve the restrictions by returning, for each word, the set of sense which do not break them (that is, those whose semantic category is at the same, or a lower, level in the hierarchy).

4 Combining Knowledge Sources

Since each of our partial taggers suggests only possible senses for each word it is necessary to have some method to combine their results. We trained decision lists (Clark and Niblett, 1989) using a supervised learning approach. Decision lists have already been successfully applied to lexical ambiguity resolution by (Yarowsky, 1995) where they perfromed well.

We present the decision list system with a number of training words for which the correct sense is known. For each of the words we supply each of its possible senses (apart from those removed from consideration by the part-of-speech filter (Section 3.2)) within a context consisting of the results from each of the partial taggers, frequency information and 10 simple collocations (first noun/verb/preposition to the left/right and first/second word to the left/right). Each sense is marked as either `appropriate` (if it is the correct sense given the context) or `inappropriate`. A learning algorithm infers a decision list which classifies senses as `appropriate` or `inappropriate` in context. The partial taggers and filters can then be run over new text and the decision list applied to the results, so as to identify the appropriate senses for words in novel contexts.

Although the decision lists are trained on a fixed vocabulary of words this does not limit the decision lists produced to those words, and our system can assign a sense to any word, provided it has a definition in LDOCE. The decision list produced consists of rules such as "if the part-of-speech is a noun and the pragmatic codes partial tagger returned a confident value for that word then that sense is appropriate for the context".

5 Producing an Evaluation Corpus

Rather than expend a vast amount of effort on manual tagging we decided to adapt two existing resources to our purposes. We took SEMCOR, a 200,000 word corpus with the content words manually tagged as part of the WordNet project. The semantic tagging was carried out under disciplined conditions using trained lexicographers with tagging inconsistencies between manual annotators controlled. SENSUS (Knight and Luk, 1994) is a large-scale ontology designed for machine-translation and was produced by merging the ontological hierarchies of WordNet and LDOCE (Bruce and Guthrie, 1992). To facilitate this merging it was necessary to derive a mapping between the senses in the two lexical resources. We used this mapping to translate the WordNet-tagged content words in SEMCOR to LDOCE tags.

The mapping is not one-to-one, and some WordNet senses are mapped onto two or three LDOCE senses when the WordNet sense does not distinguish between them. The mapping also contained significant gaps (words and senses not in the translation). SEMCOR contains 91,808 words tagged with WordNet synsets, 6,071 of which are proper names which we ignore, leaving 85,737 words which could potentially be translated. The translation contains only 36,869 words tagged with LDOCE senses, although this is a reasonable size for an evaluation corpus given this type of task (it is several orders of magnitude larger than those used by (Cowie et al., 1992)

(Harley and Glennon, 1997) (Mahesh et al., 1997)). This corpus was also constructed without the excessive cost of additional hand-tagging and does not introduce any inconsistencies which may occur with a poorly controlled tagging strategy.

6 Results

To date we have tested our system on only a portion of the text we derived from SEMCOR, which consisted of 2021 words tagged with LDOCE senses (and 12,208 words in total). The 2021 word occurances are made up from 1068 different types, with an average polysemy of 7.65. As a baseline against which to compare results we computed the percentage of words which are correctly tagged if we chose the first sense for each, which resulted in 49.8% correct disambiguation.

We trained a decision list using 1821 of the occurances (containing 1000 different types) and kept 200 (129 types) as held-back training data. When the decision list was applied to the held-back data we found 70% of the first senses correctly tagged. We also found that the system correctly identified one of the correct senses 83.4% of the time. Assuming that our tagger will perform to a similar level over all content words in our corpus if test data was avilable, and we have no evidence to the contrary, this figure equates to 92.8% correct tagging over all words in text (since, in our corpus, 42% of words tokens are ambiguous in LDOCE).

Comparative evaluation is generally difficult in word sense disambiguation due to the variation in approach and the evaluation corpora. However, it is fair to compare our work against other approaches which have attempted to disambiguate all content words in a text against some standard lexical resource, such as (Cowie et al., 1992), (Harley and Glennon, 1997), (McRoy, 1992), (Véronis and Ide, 1990) and (Mahesh et al., 1997). Neither McRoy nor Veronis & Ide provide a quantative evaluation of their system and so our performance cannot be easily compared with theirs. Mahesh et. al. claim high levels of sense tagging accuracy (about 89%), but our results are not directly comparable since its authors explicitly reject the conventional markup-training-test method used here. Cowie et. al. used LDOCE and so we can compare results using the same set of senses. Harley and Glennon used the *Cambridge International Dictionary of English* which is a comparable resource containing similar lexical information and levels of semantic distinction to LDOCE. Our result of 83% compares well with the two systems above who report 47% and 73% correct disambiguation for their most detailed level of semantic distinction. Our result is also higher than both systems at their most rough grained level of distinction (72% and 78%). These results are summarised in Table 1.

In order to compare the contribution of the separate taggers we implemented a simple voting system. By comparing the results obtained from the voting system with those from the decision list we get some idea of the advantage gained by optimising the combination of knowledge sources. The voting system provided 59% correct disambiguation, at identifying the first of the possible senses, which is little more than each knowledge source used separately (see Table 2). This provides a clear indication that there is a considerable benefit to be gained from combining disambiguation evidence in an optimal way. In future work we plan to investigate whether the apparently orthogonal, independent, sources of information are in fact so.

7 Conclusion

These experimental results show that it is possible to disambiguate all content word in a text to a high level of accuracy (92%). Our system uses an optimised combination of lexical knowledge sources which appears to be a sucessful strategyu for this problem. The results reported here are slightly lower than those for system which concentrate on small sets of words. Our future research aims to reduce this gap further.

Acknowledgments

The work described in this paper has been supported by the European Union Language Engineering project "ECRAN – Extraction of Content: Research at Near-market" (LE-2110).

References

E. Brill. 1992. A simple rule-based part of speech tagger. In *Proceeding of the Third Conference on Applied Natural Language Processing (ANLP-92)*, pages 152–155, Trento, Italy.

R. Bruce and L. Guthrie. 1992. Genus disambiguation: A study in weighted preference. In *Proceedings of COLING-92*, pages 1187–1191, Nantes, France.

P. Clark and T. Niblett. 1989. The CN2 Induction Algorithm. *Machine Learning Journal*, 3(4):261–283.

J. Cowie, L. Guthrie, and J. Guthrie. 1992. Lexical disambiguation using simulated annealing. In *Proceedings of COLING-92*, pages 359–365, Nantes, France.

W. Gale, K. Church, and D. Yarowsky. 1992. One sense per discourse. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 233–237, Harriman, NY.

A. Harley and D. Glennon. 1997. Sense tagging in action: Combining different tests with additive weights. In *Proceedings of the SIGLEX Workshop*

System	Resource	Ambiguity level	Result
(Cowie et al., 1992)	LDOCE	homograph	72%
		sense	47%
(Harley and Glennon, 1997)	CIDE	'coarse' level	78%
		'fine' level	73%
Reported system	LDOCE	sense	83%

Table 1: Comparison of tagger with similar systems

Knowledge Sources	Result
Dictionary definitions	58.1%
Pragmatic codes	55.1%
Selectional Restrictions	57%
All	59%

Table 2: Results from different knowledge sources

"Tagging Text with Lexical Semantics", pages 74–78, Washington, DC.

A. Kilgarriff. 1993. Dictionary word sense distinctions: An enquiry into their nature. *Computers and the Humanities*, 26:356–387.

K. Knight and S. Luk. 1994. Building a large knowledge base for machine tanslation. In *Proceedings of AAAI-94*, pages 185–109, Seattle, WA.

K. Mahesh, S. Nirenburg, S. Beale, E. Viegas, V. Raskin, and B. Onyshkevych. 1997. Word sense disambiguation: Why have statistics when we have these numbers? In *Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 151–159, Santa Fe, NM.

S. McRoy. 1992. Using multiple knowledge sources for word sense disambiguation. *Computational Linguistics*, 18(1):1–30.

A. Newell. 1973. Computer models of thought and language. In Schank and Colby, editors, *Artificial Intelligence and the Concept of Mind*. Freeman, San Francisco, CA.

H. Ng and H. Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of ACL-96*, pages 40–47, Santa Cruze, CA.

P. Procter, editor. 1978. *Longman Dictionary of Contemporary English*. Longman Group, Essex, UK.

H. Schütze. 1992. Dimensions of meaning. In *Proceedings of Supercomputing '92*, pages 787–796, Minneapolis, MN.

M. Stevenson. 1998. Extracting syntactic relations using heuristics. In *Proceedings of the European Summer School on Logic, Language and Information '98*, pages 248–256, Saarbrücken, Germany.

J. Véronis and N. Ide. 1990. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *Proceedings of COLING-90*, pages 389–394, Helsinki, Finland.

Y. Wilks and M. Stevenson. 1997. Combining independent knowledge sources for word sense disambiguation. In *Proceedings of the Third Conference on Recent Advances in Natural Language Processing Conference (RANLP-97)*, pages 1–7, Tzigov Chark, Bulgaria.

Y. Wilks and M. Stevenson. 1998. The grammar of sense: Using part-of-speech tags as a first step in semantic disambiguation. *Journal of Natural Language Engineering*, 4(2):135–144.

Y. Wilks, D. Fass, C.-M. Guo, J. McDonald, T. Plate, and B. Slator. 1990. A tractable machine dictionary as a basis for computational semantics. *Journal of Machine Translation*, 5:99–154.

Y. Wilks. 1997. Senses and Texts. *Computers and the Humanities*, 31:77–90.

D. Yarowsky. 1995. Unsupervised word-sense disambiguation rivaling supervised methods. In *Proceedings of ACL-95*, pages 189–196, Cambridge, MA.