Proceddings of the Conference Recent Advances in Natural Language Processing, 1997
Tzigov Chark, Bulgaria, pp1-7

Combining Independent Knowledge Sources for Word Sense Disambiguation

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

Abstract

Sense tagging, the automatic assignment of the appropriate sense from some lexicon to each of the words in a text, is a specialised instance of the general problem of word sense disambiguation. We discuss which recent word sense disambiguation algorithms are appropriate for sense tagging. It is our belief that sense tagging can be carried out effectively by combining several simple, independent, methods and we include the design of such a tagger. A prototype of this system has been implemented, correctly tagging 88% of polysemous word tokens in a small test set, providing evidence that our hypothesis is correct.

1 Sense Tagging

There has been a tendency in word sense disambiguation (WSD hereafter) literature to carry out different disambiguation tasks and classify them all as WSD procedures. There are however, at least, three different levels of algorithm in WSD.

The most specific of these are *sense tagging* procedures, these assign, to each word[1] in a text, its particular sense from some lexicon and each word type having a set of senses specific to it. This differs from the more general case of *semantic tagging*, where the tags for each word (type) need not be specific to that type and do not correspond to word senses in a lexicon. These tags may be broad semantic categories such as `HUMAN` or `ANIMATE` or WordNet synsets which may apply to any word in the text (similar to part of speech tagging in the sense that there is a class of tags which may apply to any token in the text). The most general class of algorithms are *semantic disambiguation* algorithms. These algorithms are procedures which carry out semantic disambiguation on words, these may not necessarily be tagging algorithms, in that they do not attempt to mark every token in a text but may be restricted to disambiguating small sets of word types.

The class of sense tagging algorithms is a proper subset of the class of semantic tagging algorithms, which is, in turn, a proper subset of the class of semantic disambiguation algorithms. This hierarchical relationship is represented in Figure 1.

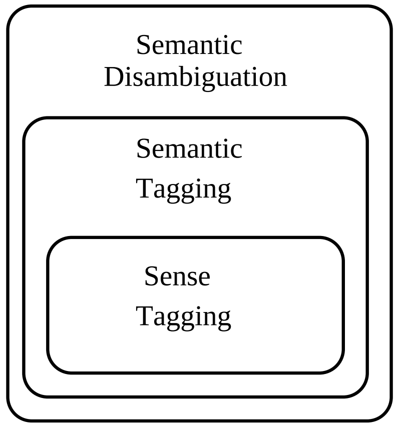

Figure 1: Levels of WSD Algorithms

In this paper we shall discuss some recent approaches to the WSD problem (Section 2) and examine their usefulness for the more specialised task of sense tagging (Section 3). We then propose an approach which makes use of several different types of information (Sections 4 and 5) and report a partial implementation of this system (Section 6) which produces very encouraging results (Section 8).

2 Recent Word Sense Disambiguation algorithms

Word sense disambiguation algorithms may be categorised by the method used to overcome the knowledge acquisition bottleneck which exists in the area. Here is one possible categorisation:

1. **Knowledge based**
WSD using information from an explicit lexicon or knowledge base. The lexicon may be a machine readable dictionary or thesaurus (such as LDOCE or WordNet) or it may be hand-crafted. Examples of this work include that of McRoy (McRoy, 1992)

2. **Corpus based**
WSD using information gained from training on some corpus. This approach can be further subclassified:

[1] This is often loosened to each *content* word.

(a) *Tagged corpora*
Information is gathered from corpora which have already been semantically disambiguated. This tagging need not be an explicit sense tag, the bilingual corpus used by Brown (Brown et al., 1991) to train a statistical WSD algorithm were disambiguated for the purposes of that approach. Training text tagged with explicit senses from a lexicon has been used by (Bruce and Wiebe, 1994) and (Ng and Lee, 1996).

(b) *Untagged corpora*
Information is gathered from raw corpora which has not been semantically disambiguated. The clustering algorithms of Pedersen and Bruce (Pedersen and Bruce, 1997) fall into this category, as does Yarowsky's algorithm to tag words with thesaurus categories (Yarowsky, 1992).

3 Comparing Different Approaches

Approach 2a is the least promising since text tagged with word senses is practically non-existent and is both time consuming and difficult to produce manually. Much of the research in this area has been compromised by the fact that researchers have focussed on lexical ambiguities that are not true word sense distinctions, such as words translated differently across two languages (Gale, Church, and Yarowsky, 1992) or homophones[2] (Yarowsky, 1993).

Even in the cases where data with the appropriate sense distinctions is available, the text is unlikely to be from the desired domain: a word sense discriminator trained on company news text will be much less effective on text about electronics products. A discriminator trained on many types of text so as to be generic will not be particularly successful in any specific domain.

Approach 2b has received much attention recently. Its disadvantage is that sense disambiguation is not carried out relative to any well defined set of senses, but rather an ad hoc set. Although this research has been the most successful of all approaches, it is difficult to see what use could be made of the word sense distinctions produced.

A recent experiment by Pedersen and Bruce (Pedersen and Bruce, 1997) attempted to avoid this problem, they used a number of clustering algorithms to infer groups of similar occurances of words in untagged text, the mapped these clusters onto senses in a lexicon, the results were compared against text which was tagged manually. Unfortumately the results were unpromising, better results could have been obtained by simply choosing the most frequently occuring sense for each word. These results force us to question the usefulness of such methods when tagging against the sorts of senses found in a general lexicon.

Approach 1 has been attempted using several different knowledge sources. The least promoising of these seems to be when the knowledge source is crafted by hand, many of the early WSD systems used this approach (for example (Wilks, 1978) and (Hirst, 1987)). An indication of the scale of the task is given by the experience of Small, who attempted WSD using "word experts" – essentially hand-crafted disambiguators, and reported that "the word expert for "throw" is "currently six pages long, but should be ten times that size" (p. 146) (Small and Reiger, 1982). The hand-crafted approach seems an impractical one for any system which aims at broad coverage.

A more practical use of this approach is to make use of one of the several lexical resources which are not available, for example WordNet (Miller (Ed.), 1990), a Machine Readable Dictionary or a Lexical Knowledge Bases (eg. (Wilks et al., 1990)).

4 Proposed Approach

Word senses are not absolute or Platonic but defined by a given lexicon, as has been known for many years from early work on WSD, even though the contrary seems widely believed: ".. it is very difficult to assign word occurrences to sense classes in any manner that is both general and determinate. In the sentences "I have a stake in this country." and "My stake in the last race was a pound" is "stake" being used in the same sense or not? If "stake" can be interpreted to mean something as vague as 'Stake as any kind of investment in any enterprise' then the answer is yes. So, if a semantic dictionary contained only two senses for "stake": that vague sense together with 'Stake as a post', then one would expect to assign the vague sense for both the sentences above. But if, on the other hand, the dictionary distinguished 'Stake as an investment' and 'Stake as an initial payment in a game or race' then the answer would be expected to be different. So, then, word sense disambiguation is relative to the dictionary of sense choices available and can have no absolute quality about it." (Wilks, 1972)

There is no general agreement over the number of senses appropriate for lexical entries: at one end of the spectrum Wierzbicka (Wierzbicka, 1989) claims words have essentially one sense while Pustejovsky believes that "... words can assume a potentially infinite number of senses in context."(Pustejovsky, 1995) How, then, are we to get an initial lexicon of word senses? We believe the best resource is still a Machine Readable Dictionary: they have a relatively well-defined set of sense tags for each word and lexical coverage is high.

We choose to use the machine readable version of the *Longman Dictionary of Contemporary English* (LDOCE) (Procter, 1978). LDOCE is a learners' dictionary – one designed not for native speakers of English but for those learning English as a second language. It has been used extensively in MRD research (e.g. (Ide and Veronis, 1994), (Cowie, Guthrie,

[2]Words pronounced identically but spelled differently.

and Guthrie, 1992), (Bruce and Wiebe, 1994)). The senses for each word in LDOCE are grouped into *homographs*, sets of senses with related meanings. For example, one homograph for "bank" contains all the senses which mean 'things piled up' (shown in Figure 2), with the different senses distinguishing exactly what has been piled.

`bank` *n*
1 land along the side of a river, lake, etc.
2 earth which is heaped up in a field or a garden, often making a border or division.
3 a mass of snow, mud, clouds, etc. *The banks of dark cloud promised a heavy storm.*
4 a slope made at bends in a road or race-track, so that they are safer for cars to go round.
5 SANDBANK: *The Dogger Bank in the North Sea can be dangerous for ships.*

Figure 2: First homograph of *bank* in LDOCE

5 Partial Taggers

An interesting fact about recent word sense disambiguation algorithms is that they have made use of different, orthogonal, sources of information: the information provided by each source seems independent of and has no bearing on any of the others. Our intuition is that word sense disambiguation can be most effectively carried out by combining these knowledge sources. We propose a disambiguation system which makes use of several partial taggers, each using one of the independent knowledge sources. None of these taggers will fully disambiguate the text but each will provide as much disambiguation information as possible given the knowledge source they use and their outputs can be combined to form the final disambigaution. Our system is in the tradition of McRoy (McRoy, 1992) although, the information sources we use are orthogonal, unlike those she used, making the evaluation of the performance of the various partial taggers more straightforward.

The partial taggers, and the knowledge sources they make use of, are listed below:

5.1 Part-of-speech filtering

It has already been shown that part-of-speech tags are a useful discriminator for semantic disambiguation (Wilks and Stevenson, 1996), although they are not, normally, enough to fully disambiguate a text. For example knowing "bank" in "My bank is on the corner." is being used as a noun will tell us that the word is not being used in the 'plane turning corner' sense but not whether it is being used in the 'financial institution' or 'edge of river' senses. Part-of-speech tags can provide a valuable step towards the solution to sense tagging: fully disambiguating about 87% of ambiguous word tokens and reducing the ambiguity for some of the rest.

5.2 Domain codes (Thesaural categories)

Pragmatic domain codes can be used to disambiguate (usually nominal) senses, as was shown by (Bruce and Guthrie, 1992) and (Yarowsky, 1992). Our intuition here is that disambiguation evidence can be gained by choosing senses which are closest in a thesaural hierarchy.

5.3 Collocates

Recent work has been done using collocations as semantic disambiguators, (Yarowsky, 1993), (Dorr, 1996), particularly for verbs. We are attempting to derive disambiguation information by examining the prepositions as given in the subcategorization frames of verbs, and in the example sentences in LDOCE.

5.4 Selectional Preferences

There has been a long tradition in NLP of using selectional preferences for WSD (Wilks, 1972). This approach has been recently used by (McRoy, 1992) and (Resnik, 1993). At its best it disambiguates both verbs, adjectives and the nouns they modify at the same time, but we shall use this information late in the disambiguation process when we hope to be reasonably confident of the senses of nouns in the text from processes such as 5.2 and 5.5.

5.5 Dictionary definitions

Lesk (Lesk, 1986) proposed a method for semantic disambiguation using the dictionary definitions of words as a measure of their semantic closeness and proposed the disambiguation of sentences by computing the overlap of definitions for a sentence. Simulated annealing, a numerical optimisation algorithm, was used to make this process practical (Cowie, Guthrie, and Guthrie, 1992), choosing an assignment of senses from as many as 10^{10} choices.

6 A Basic Tagger

We have recently implemented a basic version of this tagger, initially incorporating only the part-of-speech (5.1), dictionary definition (5.5) and domain code (5.2) partial taggers, with further stages to be added later. Our tagger consists of six modules: two which preprocess the text, allowing the remaining four to carry out the disambiguation (see Figure 3).

1. The text is first processed by a named-entity identifier, we use one developed as part of Sheffield's entry for MUC-6 (Wakao, Gaizauskas, and Humphries, 1996). This identifies certain forms of proper names in the text and classifies them as either place, person, organisation or location.

Preprocessing

- Named-entity identifier
- Dictionary look-up

Disambiguation

- Part-of-speech filter
- Simulated annealing
- Domain codes
- Scoring mechanism

Figure 3: Main modules in the tagger

We make no use of these classifications at present, however, they are of potential use to a module carrying out disambiguation using selectional restrictions.

The tagger does not attempt to disambiguate any words which are identified as part of a named-entity.

2. The remaining text is stemmed, leaving only morphological roots, and split into sentences. Then words belonging to a list of stop words[3] are removed. The words which have not been identified as part of a named entity or removed because they are not one of the stop words are considered by the system to be *ambiguous words* and are disambiguated.

   For each of the ambiguous words, its senses are extracted from LDOCE and stored with it. The textual definitions contained in the senses are processed to remove stop words and remaining words stemmed.

3. The text is tagged using the Brill tagger (Brill, 1992) and a translation is carried out using a manually defined mapping from the syntactic tags assigned by Brill (Penn Tree Bank tags (Marcus, Santorini, and Marcinkiewicz, 1993)) onto the simpler part-of-speech categories associated with LDOCE senses[4]. We then remove from consideration any of the senses whose part-of-speech is not consistent with the one assigned by the tagger, if none of the senses are consistent with the part-of-speech we assume the tagger has made an error and leave the set of senses for that word unaltered.

4. We then use the simulated annealing algorithm to optimise the dictionary definition overlap for the remaining senses. This algorithm assigns a single sense to each ambiguous word.

5. The text is then run through a module which optimises the overlap of domain codes of the senses of nouns in each paragraph in the text. The optimisation algorithm used is similar to simulated annealing, although it has been modified in two ways. Firstly, we maximise the overlap of the pragmatic codes associated with the word senses rather than the content words in their definitions. Secondly, we optimise over entire paragraphs at a time rather than just sentences, this is done because there is good evidence (Gale, Church, and Yarowsky, 1992) that a wide context, of around 100 words, is optimal when disambiguating using domain codes. This process, like the previous module, also assigns a single sense to each ambiguous word.

6. The final stage is to combine the results of the preceding processes. This is done using a very simple mechanism which we hope to improve later. For each word being considered by the tagger, we add a score of `+1` for each of the senses which are given by either the simulated annealing or domain codes modules. The sense with the highest score is chosen as the tag for that word. If there is a tie (two senses with the same score, which will happen if the two partial taggers disagree) it is broken by choosing the first sense, as listed in the dictionary. This is a sensible tie-breaker since the senses are roughly ordered by frequency of occurence in text[5].

7 Example Output

Below is an example of the senses assigned by the system for the sentence "A rapid rise in prices soon eventuated unemployment." The words "rapid", "rise", "prices", "soon", and "unemployment" are assigned LDOCE senses (of the remaining words "a" and "the" are stop words and "eventuated" was not in the dictionary and so no sense could be assigned to it). The senses assigned by our system are shown below:

- `rapid` homograph 1 sense 2 — done in a short time

[3] In our system a stop word is defined to be any word which is not a noun, verb, adjective or adverb. Prepositions are included in the list of stop words and are not disambiguated.

[4] The Brill tagger uses the tag set from the Penn Tree Bank which contains 48 tags (Marcus, Santorini, and Marcinkiewicz, 1993), LDOCE uses a set of 17 more general tags.

[5] We are using the 1st Edition of LDOCE in which the publishers make no claim that the senses are ordered by frequency of occurrence in text (although they do in later editions). However, (Guo, 1989) has found evidence that there is a correspondence between the order in which sense are listed and the frequency of occurrence.

- `rise` homograph 2 sense 1 — the act of growing greater or more powerful
- `prices` homograph 1 sense 1 — an amount of money for which a thing is offered, sold or bought
- `soon` homograph 0 sense 1 — before long; within a short time
- `unemployment` homograph 0 sense 1 — the condition of lacking a job

(The senses have additional information associated which we do not show here: domain codes, part of speech and grammatical information as well as semantic information.)

8 Evaluation

We tested our tagger on 14 hand-disambiguated sentences from the Wall Street Journal, amounting to a 314 word corpus. These texts were created for the original experiments using simulated annealing (Cowie, Guthrie, and Guthrie, 1992).

The system was tested under four different configurations, each containing a different combination of the simulated annealing and pragmatic codes partial taggers. In the first configuration neither of the partial taggers were used in this situation the scoring mechanism chooses the first sense for each word. The results of this configuration are used as a base-line against which we compare the others. The next two configurations use only one of the partial tagger each, now the scoring mechanism takes the result of the particular partial tagger as defining the sense. The final combination is one in which both partial taggers are used, the scoring mechanism combines their results in the way described above. Each configuration was tested both with and without the part-of-speech filtering partial tagger. Results for each configuration were measured to the homograph and sense level and are shown in Table 1.

It should be noticed that the addition of partial taggers improves the performance of the tagger overall in both the cases where part-of-speech filtering is used and it is turned off, although the increase when more then one tagger is combined is not equal to their combined increase over the baseline of choosing the first sense.

In the best configuration of the tagger, where the simulated annealing and pragmatic code partial taggers are used along with part-of-speech filtering, we found that 88% of words, ambiguous at the homograph level, were assigned the correct homograph and 60% the correct sense. This equates to 93% correct homograph and 76% correct sense assignment over all words.

These figures should be compared to 72% correct homograph assignment and 47% correct sense assignment, over words with more than 1 homograph, using simulated annealing alone (on the same test set). It should be noted that the granularity of sense distinctions at the LDOCE homograph level (eg. "bank" as 'edge of river' or 'financial institution') is the same as the distinctions made by current small-scale WSD algorithms (eg. (Gale, Church, and Yarowsky, 1992), (Yarowsky, 1993), (Schütze, 1992)) and our system is a true tagging algorithm, operating on free text.

Our evaluation is unsatisfactory due to the small test set, but does demonstrate that the use of independent knowledge sources leads to an improvement in the quality of disambiguation. We fully expect our results to improve with the addition of further partial taggers and the development of a more sophisticated scoring regime for the combination of their results.

9 Conclusion

In this paper we have noted that many different approaches have been taken to the word sense disambiguation problem and that several different lexical knowledge sources have been used. We hypothesised that high quality sense tagging could be achieved by combining these different knowledge sources in a single tagger.

We reported a simple sense tagger which correctly disambiguates 88% of polysemous words to the LDOCE homograph level. The tagger had access to three independent knowledge sources and it was found that the best tagging results were achieved when all three knowledge sources were used by the tagger, providing evidence that our hypothesis is true. A proposal to extend this tagger with further, mutually independent, sources of lexical information was presented.

Acknowledgements

This research was supported by the European Union Language Engineering project "ECRAN – Extraction of Content: Research at Near-market" (LE2110). We are grateful to Jim Cowie at CRL for providing the simulated annealing and scoring code and to Kevin Humphries and Hamish Cunningham of the Sheffield NLP group for advice with the implementation of the tagger.

References

Brill, E. 1992. A simple rule-based part-of-speech tagger. In *Proceeding of the Third Conference on Applied Natural Language Processing*, Trento, Italy.

Brown, P., S. Della Pietra, V. Della Pietra, and R. Mercer. 1991. Word sense disambiguation using statistical methods. In *Proceedings of the 29th Meeting of the Association for Computational Linguistics*, pages 264–270, Berkley, C.A.

Bruce, R. and L. Guthrie. 1992. Genus disambiguation: A study in weighted preference. In *Proceedings of the 14th International Conference on Com-*

		First Sense	Sim Anneal	Domain Codes	Both
without POS filter	**Homograph**	62%	72%	73%	74%
	Sense	40%	47%	48%	48%
with POS filter	**Homograph**	80%	86%	86%	88%
	Sense	54%	57%	58%	60%

Table 1: Results from different Configurations of the Tagger

putational Linguistics (COLING-92), pages 1187–1191, Nantes, France.

Bruce, R. and J. Wiebe. 1994. Word-sense disambiguation using decomposable models. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 139–145, Las Cruces, New Mexico.

Cowie, J., L. Guthrie, and J. Guthrie. 1992. Lexical disambiguation using simulated annealing. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, pages 359–365, Nantes, France.

Dorr, B. 1996. Role of word sense disambiguation in lexical aquisition: Predicting semantics from syntactic cues. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING96)*, pages 322–327, Copenhagen, Denmark.

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

Guo, C-M. 1989. Constructing a Machine Tractable Dictionary from Longman Dictionary of Contemporary English. Technical Report MCCS-89-156, Computing Research Laboratory, New Mexico State University.

Hirst, G. 1987. *Semantic Interpretation and the Resolution of Ambiguity.* CUP, Cambridge, England.

Ide, N. and J. Veronis. 1994. Have we wasted our time? In *Proceedings of the International Workshop on the Future of the Dictionary*, Grenoble.

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

Marcus, M., B. Santorini, and M. A. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Tree Bank. *Computational Linguistics*, 19(2):313–330.

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

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

Ng, H. T. and H. B. Lee. 1996. Integrating multiple knowldge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of ACL96*, pages 40–47.

Pedersen, T. and R. Bruce. 1997. Distinguishing word senses in untagged text. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, Providence, RI, August.

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

Pustejovsky, J. 1995. *The Generative Lexicon.* MIT Press, Cambridge, MA.

Resnik, P. 1993. *Selection and Information: A Class-based Approach to Lexical Relationships.* Ph.D. thesis, University of Pennsylvania.

Schütze, H. 1992. Dimensions of meaning. In *Proceedings of Supercomputing '92.*

Small, S. and C. Reiger. 1982. Parsing and comprehending with word experts (a theory and it's realiastion). In W. Lehnert and M. Ringle, editors, *Strategies for Natural Language Processing.* Lawrence Erlbaum Associates, Hillsdale, NJ.

Wakao, T., R. Gaizauskas, and K. Humphries. 1996. Evaluation of an algorithm for the recognition and calssification of proper names. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING96)*, pages 418–423, Copenhagen, Denmark.

Wierzbicka, A. 1989. *Semantics, Culture and Cognition.* Oxford University Press, Oxford.

Wilks, Y. 1972. *Grammar, Meaning and the Machine Analysis of Language.* Routledge, London.

Wilks, Y. 1978. Making preferences more active. *Artificial Intelligence*, 11(3):197–223, December.

Wilks, Y., D. Fass, CM. 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.

Wilks, Y. and M. Stevenson. 1996. The grammar of sense: Is word-sense tagging much more than part-of-speech tagging? Technical Report CS-96-05, University of Sheffield. Available as `http://xxx.lanl.gov/ps/cmp-lg/9607028`.

Yarowsky, D. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING92)*, pages 454–460, Nantes, France.

Yarowsky, D. 1993. One sense per collocation. In *Proceedings ARPA Human Language Technology Workshop*, pages 266–271, Princeton, N.J.