The Grammar of Sense: Using part-of-speech tags as a first step in semantic 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

(*Received May 1997; Revised October 1997*)

Abstract

This paper describes two experiments: one exploring the amount of information relevant to sense disambiguation contained in the part-of-speech field of entries in a Machine Readable Dictionary (MRD); the other, more practical, experiment attempts sense disambiguation of all content words in a text assigning MRD homographs as sense tags using only part-of-speech information. We have implemented a simple sense tagger which successfully tags 94% of words using this method. A plan to extend this work and implement an improved sense tagger is included.

Contents

1 Introduction

Sense tagging is the process of assigning the appropriate sense from a lexicon to each word token in a text[1], similar to the way a grammatical category is assigned in part-

[1] This is often loosened to mean only each content word in a text.

of-speech tagging. There is also a broader class of algorithms which carry out word sense disambiguation (WSD) which are not necessarily sense tagging algorithms: they may not resolve the sense of every word in a text to a unique sense or may attach semantic tags which do not correspond to word senses from some lexicon.

There is no clear consensus of opinion over whether part-of-speech disambiguation should be included as part of the WSD process. Some recent WSD algorithms (e.g. (Yar95), (Sch92)) deal only with small sets of words and treat part-of-speech ambiguity as a separate problem. While theoretically interesting, these algorithms are of limited usefulness to practical Language Engineering systems. We believe that part-of-speech ambiguity should be treated as part of the problem of word sense disambiguation, since, as we shall show, it is a lexical ambiguty problem related to that of word sense ambiguity. Since all sense tagging procedures are also word sense disambiguation algorithms, by our definitions, then sense tagging algorithms should also deal with part-of-speech ambiguity. As we have already reported (see (WFG$^+$90)) it is essential for a disambiguation algorithm to process all content words in a text if it is to be of practical use.

We have found that the resolution of part-of-speech ambiguity goes much of the way to resolving sense ambiguity and we report an interesting statistic that may cause us to reconsider how initial, rough-grained, sense tagging should be carried out.

2 Work so far

In previous work, four basic methods have been used for word sense disambiguation, corresponding to intuitions of the linkage of word-sense to:

1. collocation based methods, usually determined by the window of words in which a token occurs.
2. relevance to subject matter and surrounding text, a method first explored by Masterman in 1966 (see (WSG96)).
3. dictionary definition overlap. A method first proposed by Lesk (Les86): a sentence could be disambiguated relative to a dictionary be choosing the configuration of senses which maximises the number of words which are common to the textual definitions.

4. use of selectional constraints or preferences, as used in one of the authors' work on Preference Semantics (see (Wil75)).

Yarowsky's recent work has investigated, in seperate experiments, both methods 1 and 2 which have achieve accuracy figures of up to 96% for selected word sense distinctions. Although these methods are pioneering they are of limited usefulness to practical systems (a fact one of the authors has noted elsewhere (Wil98)), the problem with the majority of these techniques is that they are very small scale, usually disambiguation only around 10 words and assuming that these words have only two senses. Moreover, although Yarowsky has experimented with varying methods, the meaning of "sense" used varies from task to task e.g. from assigning categories from Roget's Thesaurus (Cha77) to having a bilingual Dagan/Itai-style correspondence between parallel text (DI94). One could generalise and say his results can therefore be compared to Cowie and Guthrie, though they are much superior on the smaller scale he uses, since the distinctions Yarowsky makes (e.g. Roget categories or bilingual correspondances) are equivalent to what are distinguished as homographs in LDOCE.

Method 3 has been optimised by Cowie and Guthrie (CGG92) using simulated annealing (MRR$^+$53) and they report results of 72% correct assignment of homographs from *The Longman Dictionary of Contemporary English* (LDOCE)(Pro78) and a much lower level for individual sense assignment. This result must be seen against a background figure of 62% (WSG96) correct sense assignment in LDOCE achieved by simply assigning the most frequent LDOCE sense for each word. The importance of this method is that it disambiguates all the content words in a sentence, even though it involves a vast computation for a sentence if all the LDOCE senses were considered, often optimising more than 10^9 sense combinations for a 12 word sentence.

Method 4 was one of the first to be applied to sense disambiguation (see (KF64)) and has been most recently used by CRL (MNB$^+$97) who have used it in conjunction with a massive ontology.

A key fact to notice about 1-4 is that they are methods resting on quite different intuitions: and one might well infer that, if they all capture at least part of what we intuitively mean by word-sense, then the sensible way to achieve high-quality

semantic disambiguation is to combine all four. In the next section, we shall show how we achieve high percentage, large-scale, figures using only part-of-speech information, a method different from all the above. In the conclusion we shall show how we intend to proceed by combining our current results with aspects of all the above methods to optimise our results further.

3 Experiments using part-of-speech

The hypothesis we wish to test is that part-of-speech tagging and semantic disambiguation are not as independent as has normally been assumed. Part-of-speech tagging is a well established module in many Language Engineering systems these days giving accuracy figures of up to 98% (Bri95). Our first investigation was to see how far, given a basic NLP lexicon such as a Machine Readable Dictionary, accurate part-of-speech tagging would discriminate senses without any further processing[2].

The lexicon we use is the machine readable version of LDOCE, a dictionary designed for students of English which contains around 36,000 word types. The senses for each word type are grouped into *homographs*, sets of senses with related meanings. For example, one of the homographs of "bank" means roughly 'things piled up', the different senses distinguishing exactly what is piled up (see Figure 1).[3] If the senses are sufficiently close together in meaning there will be only one homograph for that word which we call *monohomographic*. However, if the senses are far enough apart, as in the "bank" case, they will be grouped into separate homographs, which we call *polyhomographic*.

As can be seen from the example entry, each LDOCE homograph includes information about its part-of-speech (and the part-of-speech with which the homograph is marked applies to each of the senses which make uo that homograph). However, word senses are not partitioned into homographs by purely syntactic criteria: around 2% of word types in LDOCE contain a homograph which has more than one part-of-speech associated with each of its senses, which is thus a homograph with

[2] The authors are grateful to Mark Leisher of CRL at New Mexico State University who provided preliminary results on the sense distribution in a MRD, which encouraged us to conduct further research.

[3] In LDOCE "bank" has five homographs in total, we show only three for brevity.

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.*

bank *v*

1 (of a car or aircraft) to move with one side higher than the other, esp. when making a turn

bank *n*

1 a place where money is kept and paid out on demand, and where related activities go on.

2 a place where something is held ready for use, esp. organic product of human origin for medical use

3 (a person who keeps) a supply of money or pieces for payment or use in a game of chance

4 break the bank to win all the money that the BANK 4(3) as in a game of chance

Fig. 1. Some homographs of "bank" in LDOCE

multiple parts of speech. We argue later (Section 4) that homographs partitioned by grammatical categories are a natural side-effect of grouping semantically related senses.

Although the granularity of the distinction between homographs in LDOCE is rather rough-grained, they are of an appropriate level for many practical Language Engineering tasks. "bank" in the sense of 'financial institution' translates to "banque" in French but when used in the 'edge of river' sense it translates as "bord". This level of semantic disambiguation is ideal for choosing the correct target word in an English-to-French Machine Translation system and is at the same level of granularity as the sense distinctions explored by other researchers in WSD, for example (GCY92), (Sch92).

3.1 The Structure of a Lexicon: A Gedankenexperiment

We attempted to discover how useful part-of-speech information could be for semantic disambiguation. We scanned through LDOCE and examined each word type

for possible disambiguation to the homograph level by part of speech. It is possible to place each LDOCE word type in one of the following categories:

1. **Guaranteed disambiguation:** those word types for which each homograph has a distinct grammatical category associated with it.
   These words will always be disambiguated if its part-of-speech in a text is known.
   e.g. a word with 3 homographs with grammatical categories `n`, `v` and `adj`.
2. **Possible disambiguation:** those word types for which there is at least one grammatical category associated with exactly one homograph but there is another category associated with more than one.
   These words will be disambiguated by only some part-of-speech assignments, and others will not disambiguate it.
   e.g. a word whose homographs had grammatical categories `n`, `v`, `v` would be disambiguated if its part-of-speech was `n` but not if it was `v`.
3. **No disambiguation:** those for which each grammatical category that can apply to the word types is associated with more than one homograph.
   These words can never be fully disambiguated by part-of-speech alone.
   e.g. homographs with grammatical categories `v`, `v`, `n`, `n`.

The number of words which fall into the guaranteed disambiguation category puts a theoretical lower bound on the number of words which could be disambiguated if we knew the grammatical category for each word, while an upper bound may be given by adding those which could be possibly disambiguated.

We examined each word type in LDOCE (except for closed class words such as prepositions and determiners) and found that 34% were polysemous and 12% polyhomographic (a word type must, of course, be polysemous to be polyhomographic, since each homograph is a non-empty set of senses). 88% of the polyhomographic words were guaranteed to be disambiguated to the homograph level and 95% of them could possibly be disambiguated to the homograph level if their part-of-speech in a context were known. If we assume that all monohomographic words are trivially disambiguated then we can translate these values to 98.6% guaranteed disambiguation and 99.4% possible disambiguation over all word types.

This experiment of course presumes a perfect part-of-speech tagger but, as we have already mentioned, many fairly reliable taggers are readily available. It is impossible to guess how these results will translate to a real experiment since the results of this will be highly dependent upon the distribution of word types across tokens in the corpus which is being examined. In this *Gedankenexperiment* each of the word types in the dictionary is considered only once, but some word types will occur many times in a corpus and even more will never occur. So, for example, the upper bound would not apply if, by chance, none of the words of type 3 appeared in a given text.

3.2 Using a Tagger: A Practical Experiment

In order to discover what relation these theoretical results would have to the disambiguation of a real text we took five articles from the *Wall Street Journal*, containing around 1700 words in total, and disambiguated the content words using only part-of-speech tags.

The texts were part-of-speech tagged using the Brill tagger (Bri95) and the content words were flagged (the part-of-speech tags being used to decide which of the words were content words). The tags assigned by the Brill tagger were manually mapped onto the simpler part-of-speech tags used in LDOCE.[4]

Since part-of-speech tags can not assign a unique homograph to every token in text we need a method of choosing one when the tag suggests several (as may happen if the word is of the possible disambiguation or no disambiguation types). We decided to choose the first sense listed in LDOCE, as this is the one which occurs most frequently[5] and probably the best guess.

The LDOCE homographs which corresponded to the part-of-speech assigned by the tagger were extracted from the appropriate LDOCE entry and the first of those was then chosen as the sense of the word.

[4] The Brill tagger uses the tag set from the Penn Tree Bank which contains 48 tags (MSM93), 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, (Guo89) has found evidence that there is a correspondence between the order in which sense are listed and the frequency of occurrence.

We found that 92% of the content word tokens were tagged with the correct homograph compared with manual tagging of the same five texts. 57% of the content words were in fact polyhomographic and of these 87.4% were assigned the correct homograph. The monohomographic words, which made up the rest of the content words were, trivially, 100% correct. When we consider these results over all words (including non-content words) we find that 94% of all words are assigned the correct sense.

It is worth noting in passing that although only 12% of the word types in LDOCE are polyhomographic, more than half the content words in the text we tested were. This is keeping with the result reported by Zipf (Zip45) that there is a positive correlation between the number of senses a word has and its frequency of occurance in text.

4 Conclusion

Our result should not be misinterpreted as implying a reduction of semantic matters to syntactic or morphological ones and so to a loss of richness of texture in NLP. Grammatical categories are themselves essentially semantic in origin, a fact not contradicted by observing that many languages have inflectional criteria for what it is to be a particular part-of-speech. It is no answer to the question *"What is a noun in German?"* to answer that it is the part-of-speech that is regularly capitalised!

The commonsense view is that parts-of-speech is rooted in our basic ontology of the world, of How-it-is, which is a fundamentally semantic matter. In the philosophy of language this view is sometimes thought weakened by observations like those of Waismann (Wai65) that some aspects of the world (e.g. the word for light on dancing waves is his example) are captured in one language by the use of one part-of-speech but by a different part-of-speech in another, which, if true, implies that the matter cannot be semantic, in the sense of how the world is independent of ourselves and our languages. But that, fortunately for us as NLPers, is a question on which we do not need to have views: it is certainly not an issue that can divide parts-of-speech from word-sense as one that separates language from the world, or at least the perceived world.

A more persistent worry this result may exacerbate is the traditional AI view of

these matters, one shared with Bar-Hillel (BH64), that issues of word-sense were to be settled by world knowledge, not again in any objective sense, but as a function of stored codings that express the state of the world. If that view is right (and many of the authors of (SCT88) held it in 1988), then it is unacceptable that a crucial issue like word-sense be settled by matters independent of stored world knowledge.

Matters are not really so depressing, and one way to construe our current result is that low-level methods can give an effective, basic, notion of word-sense discrimination, probably close to what we are calling a homograph, and that all finer distinctions, whether one wishes to call them word-sense or not, are matters for world knowledge, which is to say, classic AI. So, one can cite simple examples such as *"He wiped the bicycle before sitting on it."* which have been used to argue that there is a sense of bicycle meaning 'bicycle seat', and then so on for each of its 250 component parts. This is plainly absurd: an extension of word-sense into an area best thought of as knowledge processing.

The current techniques can be seen as defining (at least when optimised, see Section 5) a practical limit to the extension of word-sense and thus demarcating the field of NLP from AI proper. This is perhaps no more than a sensible compromise position, consistent with the NLP discrimination methods available, namely that for Language Engineering we should take word sense disambiguation to be discrimination to the homograph level, an approach that has already proved successful (SC97).

We conclude that a useful majority of coarse-grained sense distinctions can be resolved simply by knowing a word's part-of-speech. In that small experiment using a 1700 word corpus taken from the Wall Street Journal we found that 92% of content word tokens, and 94% of all words, can be disambiguated to the LDOCE homograph using the information produced by a part-of-speech tagger. This approach is not guarenteed to solve all of the "interesting cases" of word sense disambiguation (disambiguation to the exact homograph when there are more than one with the same part-of-speech, as in our "bank" example) but it does show that a weak method can be a powerful discriminator for rough-grained semantic analysis.

However, our method is comparable to other recent work like Cowie and Guthries' simmulated annealing and Yarowsky's algorithm to tag words with their Roget cat-

egory. Like both of these our method disambiguates senses with respect to a human-constructed lexical resource of substantial size which makes use of similar levels of semantic distinction (namely the LDOCE homograph and Roget's categories).

It would be natural to question how effective our method would be when applied to a lexical resource which does not have the structural property of grouping semantically related senses, such as WordNet. However, McCarthy (McC97) compared our approach to that of Yarowsky (Yar92) and found that our simple method was more accurate as well as being computationally cheaper. It does not seem then that the success of this approach is a phenomenon ideosyncratic to one lexical resource.

5 Further work

Our next step will be to try to improve our sense tagger by using a number of weak methods and a very general, trained, scoring system to combine their results. These methods will be orthogonal in the sense that the results of any module may have little or no bearing on, and be completely independent of, all the others. This will be similar to the work done by McRoy (McR92), however we intend to carry out a strict evaluation regime, an element missing from her work, and the orthogonality of methods will make this evaluation easier. The methods we plan to use are:

1. Sense discrimination by part-of-speech
2. Subject codes (Pragmatic codes in LDOCE)
3. LDOCE example sentences as syntagmatic cues or collocates
4. Selectional restrictions or prefrences
5. Optimisation of Lesk's heuristic using simulated annealing

1 was described above. 2 has been used to produce a sense-tagged hierarchy for all LDOCE nouns at a high level of accuracy (WSG96), and is essentially the same type of information as the Roget Thesaurus used by Yarowsky in (Yar95). 3 is a limited version of the One-sense-per-collocation heuristic of Yarowsky (Yar93) which he showed had sense resolving power for almost any explicit form of collocation. We propose to use LDOCE as a source of possible signature collocates. Even if they prove a weak source of information, they will be unlikely to harm the overall sense resolution. 4 is a long established module in word sense disambiguation systems, and

we plan to use the selectional preferences in LDOCE as guides. 5 is the established Cowie and Guthrie simmulated annealing method.

Acknowledgements

This paper has been produced from research carried out under the European Union Language Engineering Project *ECRAN* (LE-2110), we are grateful for their support. It has also benefited from the comments of several members of Sheffield's NLP group, most notably Mark Hepple, Roberta Catizone and Robin Collier. We are also grateful for the comments from anonymous reviewers of this article.

References

Y. Bar-Hillel. *Language and Information.* Addison-Wesley, 1964.

E. Brill. Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics*, December 1995.

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

R. Chapman. *Roget's International Thesaurus (Fourth Edition).* Harper and Row, NY, 1977.

I. Dagan and A. Itai. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20, 1994.

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

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

J. Katz and J. Fodor. The structure of a semantic theory. In J. Katz and J. Fodor, editors, *The structure of language*, pages 479–518. Prentice Hall, NY, 1964.

M. Lesk. 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, Canada, 1986.

D. McCarthy. Word sense disambiguation for selectional restrictions. In *Proceedings of the 35th Meeting of the Association for Computational Linguistics (ACL '97)*, Madrid, 1997.

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

K. Mahesh, S. Nirenburg, S. Beale, E. Viegas, V. Raskin, and B. Onyshkeyvych. 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, July 1997.

N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equation state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

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

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

F. Segond and M. Copperman. Lexicon filtering. In *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP-3)*, pages 51–58, Tzigov Chark, Bulgaria, 1997.

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

S. Small, G. Cottrell, and M. Tanenhaus, editors. *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology and Artificial Intelligence*. Morgan Kaufmann, San Mateo, CA, 1988.

F. Waismann. *The Principles of Linguistic Philosophy*. Macmillan, London, 1965.

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

Y. Wilks. A preferential pattern-seeking semantics for natural language inference. *Artificial Intelligence*, 6:53–74, 1975.

Y. A. Wilks. Senses and texts. Submitted to (Computational Linguistics Special Issue on Word Sense Disambiguation), 1998.

Y. A. Wilks, B. M. Slator, and L. M. Guthrie. *Electric Words: Dictionaries, Computers and Meanings*. MIT Press, 1996.

D. Yarowsky. 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, 1992.

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

D. Yarowsky. Unsupervised word-sense disambiguation rivaling supervised methods. In *Proceedings of ACL95*, pages 189–196, Cambridge, MA, 1995.

G. Zipf. The meaning-frequency relationship of words. *Journal of General Psychology*, 3:251–256, 1945.