What is word sense disambiguation good for?

Adam Kilgarriff*
ITRI
University of Brighton

Abstract

Word sense disambiguation has developed as a sub-area of natural language processing, as if, like parsing, it was a well-defined task which was a prerequisite to a wide range of language-understanding applications. First, I review earlier work which shows that a set of senses for a word is only ever defined relative to a particular human purpose, and that a view of word senses as part of the linguistic furniture lacks theoretical underpinnings. Then, I investigate whether and how word sense ambiguity is in fact a problem for NLP applications.

1 What word senses are not

There is now a substantial literature on the problem of word sense disambiguation (WSD). The goal of WSD research is generally taken to be disambiguation between the senses given in a dictionary, thesaurus or similar. The idea is simple enough and could be stated as follows:

> Many words have more than one meaning. When a person understands a sentence with an ambiguous word in it, that understanding is built on the basis of just one of the meanings. So, as some part of the human language understanding process, the appropriate meaning has been chosen from the range of possibilities.

Stated in this way, it would seem that WSD might be a well-defined task, undertaken by a particular module within the human language processor. This module could then be modelled computationally in a WSD program, and this program, performing, as it did, one of the essential functions of the human language processor, would stand alongside a parser as a crucial component of a broad range of NLP applications.

There are problems with this view. The simplest stems from the observation that different dictionaries very often give different sets of senses for a word. A closer investigation reveals a lack of theoretical foundations to the concept of 'word sense'. The concept is intimately connected to our knowledge and experience of dictionaries, but these are social artifacts created to satisfy such human purposes as playing word-games, resolving family arguments, and making profits for publishers. Amid all these competing goals, the pursuit of truth is not always dominant.

In particular, a standard dictionary specifies the range of meaning of a word in a list, possibly nested, of senses. This is not the outcome of an analysis of how word-meaning operates, but is, rather, a response to constraints imposed by:

- tradition
- the printed page
- compactness
- a single, simple method of access
- resolving disputes about what a word does and does not mean.

The format of the dictionary has remained fairly stable since Dr. Johnson's day. The reasons for the format, and the reasons it has proved so resistant to change and innovation, are explored at length in Nunberg (1994). In short, the development of printed discourse, particularly the new periodicals, in England in the early part of the eighteenth century brought about a re-evaluation of the nature of meaning. No longer could it be assumed that a disagreement or confusion about a word's meaning could be settled face-to-face, and it seemed at the time that

* Research supported by the EPSRC, grant K18931.
Email: Adam.Kilgarriff@itri.bton.ac.uk

the new discourse would only be secure if there was some mutually acceptable authority on what words meant. The resolution to the crisis came in the form of Johnson's Dictionary. Thus, from its inception, the modern dictionary has had a crucial symbolic role as in-principle arbiter of disputes. Hence "the dictionary", with its implications of unique reference and authority (cf. "the Bible"). Further evidence for this position is to be found in McArthur (1987), for whom the "religious or quasi-religious tinge" (p 38) to reference materials is an enduring theme in their history; Summers (1988), whose research into dictionary use found that "settl[ing] family arguments" was a major use (p 114, cited in Béjoint (1994, p 151)); and Moon (1989) who catalogues the use of the UAD (Unidentified Authorising Dictionary) from newspapers letters pages to restaurant advertising materials (pp 60–64).

To solve disputes about meaning, a dictionary must be, above all, clear. It must draw a line around a meaning, so that a use can be classified as on one side of the line or the other. A dictionary which dwells on marginal or vague uses of a word, or which presents word meaning as context-dependent or variable or flexible, will be of little use for purposes of settling arguments. The pressure from this quarter is for the dictionary to present a set of discrete, non-overlapping meanings for a word, each defined by the necessary and sufficient conditions for its application —whatever the facts of the word's usage.

Lexicographers are vividly aware of the problem. They have frequently lamented the possibly-nested list model Stock (1983; Hanks (1994; Fillmore and Atkins (1992). They know all too well the injustice it frequently does to a word's range of meaning and use. But WSD researchers, at least until recently, have generally proceeded as if this was not the case: as if a single program — disambiguating, perhaps, in its English-language version, between the senses given in some hybrid descendant of Merriam-Webster, LDOCE, COMLEX, Roget, OALDCE and WordNet —would be relevant to a wide range of NLP applications.[1]

The sets of word senses presented in different dictionaries and thesauri have been prepared, for various purposes, for various human users: there is no *a priori* reason to believe those sets are appropriate for any NLP application.[2]

It seems likely that NLP application lexicons —which are, in the mid 1990s, almost invariably hand-built rather than MRD-derived— will be application-driven rather than resource-driven, so will only contain the word senses and make the word sense distinctions relevant to the application. They might not encounter word sense ambiguity on anything like the scale that a brief glance at a dictionary (or at the WSD literature) would suggest. The remainder of the paper addresses whether this is so, and what scale of problem word sense ambiguity causes for different varieties of NLP application.[3]

2 Taxonomy

First, let us distinguish five types of application for which WS ambiguity is potentially an issue:

- Information Retrieval (IR)
- Machine Translation (MT)
- Parsing (and, implicitly, all those applications for which parsing is one stage of processing)
- Lexicography
- Residual, 'core' language understanding (including database front ends, dialogue systems, Information Extraction as in MUC) — hereafter NLU.

2.1 IR

The intellectual affinities of most recent WSD work are with IR. The problem of finding whether a particular sense applies to an instance of a word can be construed as equivalent to the essential IR task of finding whether a document is relevant to a query. The homology is made explicit at various points in the literature (Gale et al., 1992; Gale et al., 1993).

Most work in IR disregards syntactic structure entirely, 'stemming' words so that *clean, cleaner, cleaning* and *cleaned* are all mapped to *clean*, and then treats a document as a bag of stems. It does not use POS-tagging or name-recognition, although these are relatively mature and reliable technologies for these tasks within NLP, and parsing has not been found to improve IR performance: the linguistic processing has not been fast, robust or portable enough,

[1] The most promising recent WSD work is moving away from this position, determining the senses between which the program is to disambiguate either directly from the clusters in the corpus (Schütze, 1997), or through a small amount of human input (Clear, 1994), or a choice of either (Yarowsky, 1995).

[2] For a full account of the nature of word senses, in dictionaries and elsewhere, see Kilgarriff (1992; 1993; 1997b).

[3] My sources include an informal email survey on the CORPORA mailing list, to which I had 28 responses.

and it is not in any case clear whether it provides relevant information for the IR task. This is very much a live issue: see Strzalkowski (1994), Strzalkowski and Vauthey (1995) for recent evidence of the potential of NLP in IR. However, to date, IR has made progress through applying sophisticated statistical techniques to documents treated as objects without linguistic structure, and this is the approach to WSD which has recently flourished.

Within IR, WSD can be viewed as an **alternative** to NLP, rather than a technique within it. If a statistical model based on a bag of stems is inadequate, one way to get closer to the meaning of a text is WSD; another is a linguistically-informed technique such as parsing. They are not mutually exclusive, but nor are they readily compatible.

A high proportion of WSD research is oriented towards IR, yet it is not clear whether WSD has the potential to significantly improve IR performance. In the first careful study of the question, Krovetz and Croft (1992) conducted some experiments which suggested that WS-ambiguity causes only limited degradation of IR performance. Their experiments were on the small, specialist CACM corpus. They used a standard set of queries for which "correct answers" are available. They compared system performance 'with ambiguity' and 'without ambiguity': the 'with ambiguity' condition was the normal situation, while for the 'without ambiguity' condition, all relevant terms had been manually disambiguated, in a simulation of a perfect WSD program. For this corpus and queryset, they concluded that a perfect WSD program would improve performance by 2%.

Sanderson (1994) performed a similar experiment using pseudo-words. A pseudo-word is a word formed by 'pretending' that two distinct words were a single word with two meanings, one corresponding to each of the original words. Thus the pseudo-word *banana-kalashnikov* could be formed by replacing all instances of *banana* and *kalashnikov* in a corpus by *banana-kalashnikov*: then a WSD program would have the task of determining which were originally bananas, which kalashnikovs. The method allowed Sanderson to regulate the degree of ambiguity in the corpus, and to model both accurate and inaccurate WSD programs. He found that introducing extra ambiguity did little to degrade performance, but, when the WSD algorithm made mistakes, this did do harm. Also, in longer queries the different words in the query will tend to be mutually disambiguating, so WSD is probably only relevant where the query is very short. He concludes "the performance of [IR] systems is insensitive to ambiguity but very sensitive to erroneous disambiguation" (p 149).

Schütze (1997) first distinguishes sense *discrimination* from disambiguation. Discrimination involves identifying distinct senses and classifying occurrences of the word as belonging to one of those senses. It does not involve labelling the senses (which correspond to clusters of occurrences) or associating them with any external knowledge source such as a dictionary. Thus, in keeping with the spirit of this paper, his senses are automatically devised to match the corpus. System performance improved by up to 4.3%.[4] with the addition of the disambiguation module (and the added sophistication that a word can be assigned to more than one word sense, where it is 'near' more than one in vector space).

It is debatable how important an improvement of 2 or 4 percentage points is. On the one hand, WSD will clearly not revolutionise IR or render it a solved problem. But IR is a fairly mature technology, very widely used by millions of users, and an average 4% improvement across all those users and all their many queries could be seen as very significant indeed.

2.2 Machine Translation

In IR, it is generally difficult to assign blame for poor performance to word sense ambiguity or any other specific source. MT, by contrast, wears its mistakes on its sleeve. It is abundantly clear to all in MT that word sense ambiguity is a huge problem.

The literature has surprisingly little to say about it. Hutchins and Somers (1992) point out the two variants of the problem: monolingual ambiguity, where the word is ambiguous in the source language, and translational ambiguity, where speakers of the source language do not consider the word ambiguous but it has two possible translations, as when English *blue* is translated differently into Russian according to whether it is light blue or dark.

MT is a technology rather than a science. MT systems generally take a decade from idea to marketplace, so the theory available at their inception is destined to be out of date by the time they perform. Thus no recent WSD work is employed in existing MT systems. They use extensive sets

[4] They cite an improved average precision (over 11 levels of recall) of 14.4% compared to the baseline, from 29.9% to 34.2%. This improvement is 4.3% in absolute terms, but 14.4% when calculated as an improvement on the baseline performance.

of selection restrictions paired with semantic features to make it possible for the system to make the correct lexical choice. MT systems usually use a number of very large lexicons where selection restriction information, designed to resolve ambiguity problems, accounts for a large proportion of the bulk. The SYSTRAN English-French lexicon responsible for word choice contains 400 rules governing the one English word, *oil*, and when it should be translated as *huile*, when *pétrole* (Hutchins and Somers, 1992, p 179).

One paper which does bring state-of-the-art WSD to bear on Machine Translation, albeit in experimental mode, is Dagan and Itai (1994). They use a bilingual lexicon to identify the possible translations, and a parsed target language corpus to gather information about the 'tuples' in which each of the possible translations is often found. A 'tuple' comprises a grammatical relation, such as SUBJECT-VERB, and the occupier of each of the slots of that relation, so "The man walked home" would give the triple (SUBJECT-VERB, man, walk). The source-language text to be translated in then parsed, to give a source language tuple. The bilingual dictionary and the target-language statistics are then used to find the best match.

The paper applies sophisticated WSD to a real problem, with the discriminations that the system makes being defined by the needs of the application.

2.3 Parsing

Accurate parsing is a requirement for a wide range of NLP applications, so if WSD is critical for parsing accurately, it is, by implication, significant for all those applications that depend on parsing. McCarthy (1997) explores WSD methods explicitly for purposes of improving parsing. Before assessing whether WS ambiguity is critical, let us take a step back.

It is well-established that "the problem of syntactic ambiguity is AI-complete" (Hobbs et al., 1992, p 269). Here, let us focus on one particular, but pervasive, variety of syntactic ambiguity: prepositional phrase (PP) attachment. A problem is AI-complete if its solution requires a solution to all the general AI problems of representing and reasoning about arbitrary real-world knowledge. In principle, any item of general knowledge might be the datum required to make a PP-attachment. If that is all that can be said, the outlook is bleak. We would hope that, in practice, a small and tractable subset of general knowledge will resolve a high proportion of ambiguities.

Some approaches to high-quality parsing make extensive use of machine-readable dictionaries (MRDs). In the 1990s, Microsoft have been the leading proponents of 'MRDs-for-parsing'.[5] The hypothesis behind the approach is that dictionary entries provide, implicitly or explicitly, the information required to resolve most syntactic ambiguities.

Note that, even if this hypothesis is true, it does not imply that WSD has an important role to play. Lexical information can resolve many syntactic ambiguities without being sense-disambiguated. Consider

1 I love baking cakes with friends.

2 I love baking cakes with butter icing.

The PP attachment ambiguity is resolved, along with the ambiguity of *with*, by the semantic class of the final noun phrase. Where the head of this noun phrase is human, as in 1, the PP attaches to the verb. Where it is a cake ingredient, it attaches to *cakes*. Lexical information is required to determine the attachment in 1 and 2, but, since neither *friends* nor *icing* is ambiguous between humans and cake-ingredients, disambiguation is not required.

That lexical information will resolve a high proportion of syntactic ambiguities is one hypothesis; that a significantly higher proportion will be resolved, if the lexical information is sense-specific, is another.

Almost no work has been done to test either hypothesis. Whittemore et al. (1990) tested and confirmed a related hypothesis: that 'lexical preferences' of nouns and verbs for PPs of a particular type are better predictors of PP-attachment than any purely syntactic considerations. They took a sample corpus and counted the PPs that would be correctly attached if each strategy was used. To discover the significance of WS-ambiguity to parsing, a study is required which combines this method with Krovetz and Croft's, of manually disambiguating to determine the performance improvement that would be achieved with a perfect WSD program.

2.4 Lexicography

NLP is most aware of lexicographers as suppliers of wares, but they are also customers. A linguistically annotated corpus is of more use to a lexicographer than a 'raw' one, as he or she can then investigate the behaviour of a word in particular linguistic contexts without having to

[5]The method is used in the parser embedded in 1997 Microsoft Word's grammar checker, as demonstrated by Steve Richardson at the ACL Conference in Applied NLP, Washington D.C., 1997.

trawl through large numbers of irrelevant citations. A sense-annotated corpus would be particularly valuable, as the lexicographer would not have to trawl through 'money *bank*' citations when defining 'river *bank*' (Clear, 1994). There is then an intriguing possibility that the behaviour of WSD programs will feed back into the nature of the dictionary senses they disambiguate between.

2.5 NLU

For existing NLP applications requiring a deeper understanding of the text, 99% of the ambiguity to be found in a desk dictionary is not relevant. This is, firstly, because these applications deal only with very specific text types. The specific sublanguage generally means that, if a word has a meaning which is of interest, it is very likely that occurrences of the word will be being used in that meaning and not some other. Secondly, even then the application can only interpret those inputs for which there is a possible interpretation in the knowledge base (or in the system's output behaviour). Several respondents to the email survey, where I asked, "does WS ambiguity cause problems for your system?", commented "We don't have any semantics in our lexicon, we just have hooks into the knowledge representation".

Where a word has one sense in the domain model, and one or more outside it, an NLU application can generally determine whether the word is being used in the domain sense by identifying whether the entire sentence or query is coherent in terms of the domain model. If it is, the word is almost certainly being used in the domain sense. Where a word has more than one domain sense, it is unlikely that both will produce coherent analyses. The domain model will generally provide disambiguating material, not because it has been explicitly added, but because type-checking and coherence-checking which is necessary in any case will reject invalid senses.

With time, NLU systems will become more sophisticated, with richer domain models and less limitations in the varieties of text they can analyse. This will make WSD more salient, though different strategies will be relevant for the 'foreground lexicon' containing the key words for the domain model, and the 'background lexicon', containing all other words. Foreground lexicon senses will be tightly-defined and domain-specific, and will be disambiguated by coherence-checking. Background lexicon disambiguation will only need to be between coarse-grained senses. Its function will be to increase parse accuracy, and statistical methods will be appropriate. (The full argument is presented in Kilgarriff (1997a).)

3 Answers

The answers to the question, "Does WS ambiguity cause problems for NLP applications?" are:

IR: yes, to some moderate degree. Problems can substantially be overcome by using longer queries. Within IR, WSD features as something of an alternative to NLP.

MT: yes. Huge problem, with the problem space defined by all the one-to-many and many-to-many mappings in a bilingual dictionary. Addressed to date by lots and lots of selection restrictions.

Parsing: not known.

Lexicography: yes, WSD would be of benefit.

NLU: not much. NLU applications are mostly domain specific, and have some sort of domain model. It is generally necessary to have a detailed knowledge of the word senses that are in the domain, so the knowledge to disambiguate will often be available in the domain model even where it has not explicitly been added for disambiguation purposes.

References

Henri Béjoint. 1994. *Tradition and Innovation in Modern English Dictionaries.* OUP, Oxford.

Jeremy Clear. 1994. I can't see the sense in a large corpus. In Ferenc Kiefer, Gabor Kiss, and Julia Pajzs, editors, *Papers in Computational Lexicography: COMPLEX '94*, pages 33–48, Budapest.

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

Charles J. Fillmore and Beryl T. S. Atkins. 1992. Towards a frame-based lexicon: the semantics of RISK and its neighbours. In Adrienne Lehrer and Eva Kittay, editors, *Frames, Fields and Contrasts*, pages 75–102. Lawrence Erlbaum, New Jersey.

William Gale, Kenneth Church, and David Yarowsky. 1992. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings, 30th ACL*, pages 249–156.

William Gale, Kenneth Church, and David Yarowsky. 1993. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26(1–2):415–439.

Patrick Hanks. 1994. Linguistic norms and pragmatic exploitations or, why lexicographers need prototype theory, and vice versa. In Ferenc Kiefer, Gabor Kiss, and Julia Pajzs, editors, *Papers in Computational Lexicography: COMPLEX '94*, pages 89–113, Budapest.

Jerry R. Hobbs, Douglas Appelt, Mabry Tyson, John Bear, and David Israel. 1992. Description of the FASTUS system used for MUC-4. In *Proceedings, 4th Message Understanding Conference*, pages 268–275.

John Hutchins and Harold Somers. 1992. *Introduction to Machine Translation.* Academic Press.

Adam Kilgarriff. 1992. *Polysemy.* Ph.D. thesis, University of Sussex, CSRP 261, School of Cognitive and Computing Sciences.

Adam Kilgarriff. 1993. Dictionary word sense distinctions: An enquiry into their nature. *Computers and the Humanities*, 26(1–2):365–387.

Adam Kilgarriff. 1997a. Foreground and background lexicons and word sense disambiguation for information extraction. In *Proc. Workshop on Lexicon Driven Information Extraction*, Frascati, Italy, July.

Adam Kilgarriff. 1997b. 'I don't believe in word senses'. *Computers and the Humanities*, forthcoming.

Robert Krovetz and W. Bruce Croft. 1992. Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems*, 10(2):115–141.

Tom McArthur. 1987. *Worlds of reference.* CUP, Cambridge, England.

Diana McCarthy. 1997. Word sense disambiguation for acquisition of selectional preferences. In *Proc. ACL/EACL Workshop on Automatic Information Extraction and building of lexical semantic resources*, pages 52–60, Madrid, July. ACL.

Rosamund Moon. 1989. Objective or objectionable? ideological aspects of dictionaries. *English Language Research*, 3: Language and Ideology:59–94.

Geoffrey Nunberg. 1994. The once and future dictionary. Presentation at *The Future of the Dictionary* Workshop, Uriage-les-Bains, France, October.

Mark Sanderson. 1994. Word sense disambiguation and information retrieval. In *Proceedings, ACM Special Interest Group on Information retrieval*, pages 142–151.

Hinrich Schütze. 1997. Automatic word sense discrimination. *Computational Linguistics*, forthcoming.

Penelope F. Stock. 1983. Polysemy. In *Proc. Exeter Lexicography Conference*, pages 131–140.

Tomek Strzalkowski and Barbara Vauthey. 1995. Information retrieval using robust natural language processing. In *AAAI Spring Symposium on Representation and Acquisition of Lexical Information*, pages 104–111, Stanford.

Tomek Strzalkowski. 1994. Robust text processing in automated inofrmation retrieval. In *4th Conference on Applied Natural Language Processing*, pages 168–173, Stuttgart, October.

Della Summers. 1988. The role of dictionaries in language learning. In R. A. Carter and M. McCarthy, editors, *Vocabulary and Language Teaching*, pages 111–125. Longman, London.

Greg Whittemore, Kathleen Ferrara, and Hans Brunner. 1990. Empirical study of predictive powers of simple attachment schemes for post-modifier prepositional phrases. In *ACL Proceedings, 28th Annual Meeting*, pages 23–30, Pittsburgh.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivalling supervised methods. In *ACL 95*, pages 189–196, MIT.