Senses and Texts

Yorick Wilks
Department of Computer Science
University of Sheffield,
211 Portobello Street,
Sheffield, S1 4DP, UK.
yorick@dcs.sheffield.ac.uk

Abstract

This paper addresses the question of whether it is possible to sense-tag systematically, and on a large scale, and how we should assess progress so far. That is to say, how to attach each occurrence of a word in a text to one and only one sense in a dictionary---a particular dictionary of course, and that is part of the problem. The paper does not propose a solution to the question, though we have reported empirical findings elsewhere (Cowie et al. 1992 and Wilks et al. 1996), and intend to continue and refine that work. The point of this paper is to examine two well-known contributions critically, one (Kilgarriff 1993) which is widely taken as showing that the task, as defined, cannot be carried out systematically by humans, and secondly (Yarowsky 1995) which claims strikingly good results at doing exactly that.

Introduction

Empirical, corpus-based, computational linguistics reached by now into almost every crevice of the subject, and perhaps pragmatics will soon succumb. Semantics, if we may assume the sense-tagging task is semantic, taken broadly, has shown striking progress in the last five years and, in Yarowsky's most recent work (1995) has produced very high levels of success in the 90%s, well above the key bench-mark figure of 62% correct sense assignment, achieved at an informal experiment in New Mexico about 1990, in which each word was assigned its FIRST sense listed in LDOCE.

A crucial question in this paper will be whether recent work in sense-tagging has in fact given us the breakthrough in scale that is now obvious with, say, part-of-speech tagging. Our conclusion will be that it has not, and that the experiments so far, however high their success rates, are not yet of a scale different from those of the previous generation of linguistic, symbolic-AI or connectionist approaches to the very same problem.

A historian of our field might glance back at this point to, say, Small et al. (1988) which covered the AI-symbolic and connectionist traditions of sense-tagging at just the moment before corpus-driven empirical methods began to revive. All the key issues still unsettled are discussed there and the collection showed no naivet there about the problem of sense resolution with respect only to existing lexicons of senses. It was realised that that task was only meaningful against an assumption of some method for capturing new (new to the chosen lexicon, that is) senses and, most importantly, that although existing lexicons differed, they did not differ arbitrarily much. The book also demonstrated that there was also strong psychological backing for the reality of word senses and for empirical methods of locating them from corpora without any prior assumptions about their number or distribution (e.g. Plate's work in Wilks et al. 1990, and see also Jorgensen, 1990).

Our purpose in this paper will be to argue that Kilgarriff's negative claims are simply wrong, and his errors must be combated, while Yarowsky is largely right although we have some queries about the details and the interpretation of his claims. Both authors however agree that this is a traditional and important task: one often cited as being, because of the inability of systems of the past to carry it out, a foundational lacuna in, say, the history of machine translation (MT). It was assumed by many, in that distant period, that if only word-sense ambiguity could be tamed, by the process we are calling sense-tagging, then MT of high quality would be relatively straightforward. Like may linguistic tasks, it became an end in itself, like syntactic parsing, and , now that it is, we would claim, firmly in sight (despite Kilgarriff) it is far less clear that its solution will automatically solve a range of traditional problems like MT. But clearly it would be a generally good tool to have and local triumph if this long-resistant bastion of NLP were to yield.

The very possibility of sense-tagging

Kilgarriff's paper (1993) is important because it has been widely cited as showing that the senses of a word, as distinguished in a dictionary such as LDOCE, do not cover the senses actually carried by most occurrences of the word as they appear in a corpus. If his paper does show that, it is very significant indeed, because that would imply that sense-tagging word occurrences in a corpus by means of any lexical data based on, or related to, a machine-readable dictionary or thesaurus is misguided. I want to show that here the paper does not demonstrate any such thing. Moreover, it proceeds by means of a straw-man it may be worth bringing back to life!

That straw-man, Kilgarriff's starting point, is the 'bank model' (BM) of lexical ambiguity resolution, which is established by assertion rather than quotation, though it is attributed to Small, Hirst, and Cottrell as well as the present author. In the BM, words have discrete meanings, and the human reader (like the ideal computer program) knows instantly and effortlessly which meaning of the word applies (Ibid. p.367), "given that a word occurrence always refers to one or the other, but not both" of the pair of main meanings that a word like 'bank' is reputed to have. The main thrust of Kilgarriff's paper is to distinguish a number of relationships between LDOCE senses that are not discrete in that way, and then to go on to an experiment with a corpus.

But first we should breathe a little life back into the BM straw-man: those named above can look after themselves, but here is a passage from Wilks (1972, p.12) "..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 on 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". QED.

In general, it is probably wise to believe, even if it is not always true, that authors in the past were no more naive than those now working, and were probably writing programs, however primitive and ineffective, to carry out the very same tasks as now (e.g. sense-tagging of corpus words). More importantly, the work quoted, which became an approach called preference semantics, was essentially a study of the divergence of corpus usage from lexical norms (or preferences) and developed in the Seventies into a set of processes for accommodating divergent/non-standard/metaphorical or what-you-will usage to existing lexical norms, notions that Kilgarriff seems to believe only developed in a much later and smarter group of people around 1990, which includes himself, but also, for example, Fass whose work was a direct continuation of that quoted above. Indeed, in Wilks (1972) procedures were programmed (and run over a set of newspaper editorials) to accommodate the divergent usage to that of an established sense of another word in the same text, while in Wilks (1978) programmed procedures were specified to accommodate such usage by constructing completely new sense entries.

A much more significant omission, one that bears directly on his main claim and is not merely an issue of historical correctness, is the lack of reference to work in New Mexico and elsewhere (e.g. Cowie et al. 1992) on the large-scale sense tagging of corpora against an MRD-derived lexical data base. These were larger scale experiments whose results directly contradict the result he is believed to have proved. I shall return to this point in a moment. The best part of Kilgarriff's paper is his attempt to give an intuitive account of developmental relations between the senses of a word: there is, of course, a large scholarly literature on this. He distinguishes Generalizing Metaphors (a move from a specific case to a more general one), from Must-be-theres (the applicability of one sense requires the applicability of another, as when an act of matricide requires there to be a mother); from Domain shift (where a sense in one domain, like "mellow" of wine, is far enough from the domain of "mellow" of a personality, to constitute a sense shift).

It is not always easy to distinguish the first two types, since both rest on an implication relationship between two or more senses. Again, the details do not matter: what he has shown convincingly is that, as in the earlier quotation, the choice between senses of a given word is often not easy to make because it depends on their relationship, the nature of the definitions and how specific they are. I suspect no one has ever held a simple-minded version of the BM, except possibly Fodor and Katz, who, whatever their virtues, had no interest at all in lexicography.

The real problem with Kilgarriff's analysis of sense types is that he conflates:

a) text usage different from that shown in a whole list of stored senses for a given word e.g. in a dictionary, (which is what his later experiment will be about) with

b) text usage divergent from some "core" sense in the lexicon.

Only the second is properly in the area of metaphor/metonymy or "grinding" (Copestake and Briscoe, 1991) work of the group in which he places himself, and it is this phenomenon to which his classification of sense distinctions summarized above properly belongs. This notion requires some idea of sense development; of senses of a word extending in time in a non-random manner, and is a linguistic tradition of analysis going back to Givon (1967). However, the straw-man BM and the experiment he then does on hand-tagging of senses in text, all attach to the first, unrelated, notion which does not normally imply the presence of metonymy or metaphor at all, but simply an inadequate sense list. Of course, the two types may be historically related, in that some of the (a) list may have been derived by metaphorical/metonymic processes from a (b) word, but this is not be so in general. This confusion of targets is a weakness in the paper, since it makes it difficult to be sure what he wants us to conclude from the experiment. However, since we shall show his results are not valid, this distinction may not matter too much.

One might add here that Kilgarriff's pessimism has gone hand in hand with some very interesting surveys he has conducted over the Internet on the real need for word-sense disambiguation by NLP R&D. And one should note that there are others (e.g. Ide and Veronis, 1994) who have questioned the practical usefulness of data derived at many sites from MRDs. Our case here, of course, is that it has been useful, both in our own work on sense-tagging (Cowie et al.op.cit.) and in that of Yarowsky, using Roget and discussed below.

Kilgarriff's experiment, which what has been widely taken to be the main message of his paper, is not described in much detail. In a footnote, he refuses to give the reader the statistics on which his result was based even though the text quite clearly contains a claim (p. 378) that 87% of (non-monsemous) words in his text sample have at least one text occurrence that cannot be associated with one and only one LDOCE sense. Hence, he claims, poor old BM is refuted, yet again.

But that claim (about word types) is wholly consistent with, for example, 99% of text usage (of word tokens) being associated with one and only one dictionary sense! Thus the actual claim in the paper is not at all what it has been taken to show, and is highly misleading.

But much empirical evidence tells also against the claim Kilgarriff is believed to have made. Informal analyses (1989) by Georgia Green suggested that some 20% of text usage (i.e. to word tokens) could not be associated with a unique dictionary sense. Consistent with that, too, is the use of simulated annealing techniques by Cowie et al. (1992) at CRL-New Mexico to assign LDOCE senses to a corpus. In that work, it was shown that about 75%-80% of word usage could be correctly associated with LDOCE senses, as compared with hand-tagged control text. It was, and still is, hoped that that figure can be raised by additional filtering techniques.

The two considerations above show, from quite different sources and techniques, the dubious nature of Kilgarriff's claim. Wierzbicka (1989 following Antal 1963) has long argued that words have only core senses and that dictionaries/lexicons should express that single sense and leave all further sense refinement to some other process, such as real world knowledge manipulations, AI if you wish, but not a process that uses the lexicon. Since the CRL result suggested that the automatic procedures worked very well (nearer 80%) at the homograph, rather than the sub-sense, level (the latter being where Kilgarriff's examples all lie) one possible way forward for NLP would be to go some of the way with Wierzbicka's views and restrict lexical sense distinctions to the homograph level. Then sense tagging could perhaps be done at the success level of part-of speech tagging. Such a move could be seen as changing the data to suit what you can accomplish, or as reinstating AI and pragmatics within NLP for the kind of endless, context-driven, inferences we need in real situations.

This suggestion is rather different from Kilgarriff's conclusion: which is also an empirical one. He proposes that the real basis of sense distinction be established by usage clustering techniques applied to corpora. This is an excellent idea and recent work at IBM (Brown et al. 1991) has produced striking non-seeded clusters of corpus usages, many of them displaying a similarity close to an intuitive notion of sense.

But there are serious problems in moving any kind of lexicography, traditional or computational, onto any such basis. Hanks (1994) has claimed that a dictionary could be written that consisted entirely of usages, and has investigated how those might be clustered for purely lexicographic purposes, yet it remains unclear what kind of volume could result from such a project or who would buy it and how they could use it. One way to think of such a product would be the reduction of monolingual dictionaries to thesauri, so that to look up a word becomes to look up which row or rows of context bound semi-synonyms it appears in. Thesauri have a real function both for native and non-native speakers of a language, but they rely on the reader knowing what some or all of the words in a row or class mean because they give no explanations. To reduce word sense separation to synonym classes, without explanations attached would limit a dictionary's use in a striking way.

If we then think not of dictionaries for human use but NLP lexicons, the situation might seem more welcoming for Kilgarriff's suggestion, since he could be seen as suggesting, say, a new version of WordNet (Miller, 1985) with its synsets established not a priori but by statistical corpus clustering. This is indeed a notion that has been kicked around in NLP for a while and is probably worth a try. There are still difficulties: first, that any such clustering process produces not only the clean, neat, classes like IBM's (Hindu Jew Christian Bhuddist) example but inevitable monsters, produced by some quirk of a particular corpus. Those could, of course, be hand weeded but that is not an automatic process.

Secondly, as is also well known, what classes you get, or rather, the generality of the classes you get, depends on parameter settings in the clustering algorithm: those obtained at different settings may or may not correspond nicely to, say, different levels of a standard lexical hierarchy. They probably will not, since hierarchies are discrete in terms of levels and the parameters used are continuous but, even when they do, there will be none of the hierarchical terms attached, of the sort available in WordNet (e.g. ANIMAL or DOMESTIC ANIMAL). And this is only a special case of the general problem of clustering algorithms, well known in information retrieval, that the clusters so found do not come with names or features attached.

Thirdly, and this may be the most significant point for Kilgarriff's proposal, there will always be some match of such empirical clusters to any new text occurrence of a word and, to that degree, sense-tagging in text is bound to succeed by such a methodology, given the origin of the clusters and the fact that a closest match to one of a set of clusters can always be found. The problem is how you interpret that result because, in this methodology, no hand-tagged text will be available as a control since it is not clear what task the human controls could be asked to carry out. Subjects may find traditional sense-tagging (against e.g. LDOCE senses) hard but it is a comprehensible task, because of the role dictionaries and their associated senses have in our cultural world. But the new task (attach one and only one of the classes in which the word appears to its use at this point) is rather less well defined. But again, a range of original and ingenious suggestions may make this task much more tractable, and senses so tagged (against WordNet style classes, though empirically derived) could certainly assist real tasks like MT even if they did not turn out wholly original dictionaries for the book buying public.

There is, of course, no contradiction between, on the one hand, my suggestion for a compaction of lexicons towards core or homograph senses, done to optimize the sense-tagging process and, on the other, his suggestion for an empirical basis for the establishment of synsets, or clusters that constitute senses. Given that there are problems with wholly empirically-based sense clusters of the sort mentioned above, the natural move would be to suggest some form of hybrid derivation from corpus statistics, taken together with some machine-readable source of synsets: WordNet itself, standard thesauri, and even bilingual dictionaries which are also convenient reductions of a language to word sets grouped by sense (normally by reference to a word in another language, of course). As many have now realised, both the pure corpus methods and the large-scale hand-crafted sources have their virtues, and their own particular systematic errors, and the hope has to be that clever procedures can cause those to cancel, rather than reinforce, each other. But all that is future work, and beyond the scope of a critical note.

In conclusion, it may be worth noting that the BM, in some form, is probably inescapable, at least in the form of what Pustejovsky (1995) calls a "sense enumerative lexicon", and against which he inveighs for some twenty pages before going on to use one for his illustrations, as we all do, including all lexicographers. This is not hypocrisy but a confusion close to that between (a) and (b) above: we, as language users and computational modellers, must be able, now or later, to capture a usage that differs from some established sense (problem b above), but that is only loosely connected to problem (a), where senses, if they are real, seem to come in lists and it is with them we must sense-tag if the task is to be possible at all.

Recent experiments in sense-tagging

We now turn to the claims in (Gale, Church & Yarowsky 1992, abbreviated to GCY, see also Yarowsky 1991, 1993 and 1995) that:

(1) That word tokens in text tend to occur with a smaller number of senses than often supposed and, most specifically,

(2) In a single discourse a word will appear in one and only one sense, even if several are listed for it in a lexicon, at a level of about 94% likelihood for non-monosemous words (a figure that naturally becomes higher if the monosemous text words are added in).

These are most important claims if true for they would, at a stroke, remove a major excuse for the bad progress of MT; make redundant a whole sub-industry of NLP, namely sense resolution, and greatly simplify the currently fashionable NLP task of sense-tagging texts by any method whatever (e.g. Cowie et al. op cit., Bruce & Wiebe 1994).

GCY's claim would not make sense-tagging of text irrelevant, of course, for it would only allow one to assume that resolving any single token of a word (by any method at all) in a text would then serve for all occurrences in the text, at a high level of probability. Or, one could amalgamate all contexts for a word and resolve those taken together to some pre-established lexical sense. Naturally, these procedures would be absurd if one were not already convinced of the truth of the claim.

GCY's claims are not directly related to those of Kilgarriff, who aimed to show only that it was difficult to assign text tokens to any lexical sense at all. Indeed, Kilgarriff and GCY use quite different procedures: Kilgarriff's is one of assigning a word token in context to one of a set of lexical sense descriptions, while GCY's is one of assessing whether or not two tokens in context are the same sense or not. The procedures are incommensurable and no outcome on one would be predictive for the other: GCYs procedures do not use standard lexicons and are in terms of closeness-of-fit, which means that, unlike Kilgarriff's, they can never fail to match a text token to a sense, defined in the way they do (see below).

However, GCYs claims are incompatible with Kilgarriff's in spirit in that Kilgarriff assumes there is a lot of polysemy about and that resolving it is tricky, where GCY assume the opposite.

Both Kilgarriff and GCY have given rise to potent myths about word-sense tagging in text that I believe are wrong, or at best unproven. Kilgarriff's paper, as we saw earlier, has some subtle analysis but one crucial statistical flaw. GCY's is quite different: it is a mush of hard to interpret claims and procedures, but ones that may still, nonetheless, be basically true.

GCY's methodology is essentially impressionistic: the texts they chose are, of course, those available, which turn out to be Grolier's Encyclopaedia. There is no dispute about one-sense-per-discourse (their name for claim (2) above) for certain classes of texts: the more technical a text the more anyone, whatever their other prejudices about language, would expect the claim to be true. Announcing that the claim had been shown true for mathematical or chemical texts would surprise no one; encyclopaedias are also technical texts.

Their key fact in support of claim (1) above, based on a sense-tagging of 97 selected word types in the whole Encyclopaedia, and sense tagged by the statistical method described below, was that 7569 of the tokens associated with those types are monosemous in the corpus, while 6725 are of words with more than two senses. Curiously, they claim this shows "most words (both by token and by type) have only one sense". I have no idea whether to be surprised by this figure or not but it certainly does nothing to show that (op.cit., 1992) "Perhaps word sense disambiguation is not as difficult as we might have thought". It shows me that, even in fairly technical prose like that of an encyclopaedia, nearly half the words occur in more than one sense.

And that fact, of course, has no relation at all to mono- or poly-semousness in whatever base lexicon we happen to be using in an NLP system. Given a large lexicon, based on say the OED, one could safely assume that virtually all words are polysemous. As will be often the case, GCY's claim at this point is true of exactly the domain they are dealing with, and their (non-stated) assumption that any lexicon is created for the domain text they are dealing with and with no relation to any other lexicon for any other text. One claim per discourse, one might say.

This last point is fundamental because we know that distinctions of sense are lexicon- or procedure-dependent. Kilgarriff faced this explicitly, and took LDOCE as an admittedly arbitrary starting point. GCY never discuss the issue, which makes all their claims about numbers of senses totally, but inexplicitly, dependent on the procedures they have adopted in their experiments to give a canonical sense-tagging against which to test their claims.

This is a real problem for them. They admit right away that few or no extensive hand-tagged sense-resolved corpora exist for control purposes, So, they must adopt a sense-discrimination procedure to provide their data that is unsupervised. This is where the ingenuity of the paper comes in, but also its fragility. They have two methods for providing sense-tagged data against which to test their one-sense-per-discourse claim (2).

The first rests on a criterion of sense distinction provided by correspondence to differing non-English words in a parallel corpus, in their case the French-English Canadian Hansard because, as always, it is there!. So, the correspondence of "duty" to an aligned sentence containing either "devoir" or "impot" (i.e. obligation or tax) is taken as an effective method of distinguishing the obligation/tax senses of the English word, which was indeed the criterion for sense argued for in (Dagan and Itai, 1994). It has well known drawbacks: most obviously that whatever we mean by sense distinction in English, it is unlikely to be criterially revealed by what the French happen to do in their language.

More relevantly to the particular case, GCY found it very hard to find plausible pairs for test, which must not of course SHARE ambiguities across the French/English boundaries (as interest/interet do). In the end they were reduced to a test based on the six (!) pairs they found in the Hansard corpus that met their criteria for sense separation and occurrence more than 150 times in two or more senses. In GCYs defence one could argue that, since they do not expect much polysemy in texts, examples of this sort would, of course, be hard to find.

Taking this bilingual method of sense-tagging for the six word set as criterial they then run their basic word sense discrimination method over the English Hansard data. This consists, very roughly, of a training method over 100 word surrounding contexts for 60 instances of each member of a pair of senses (hand selected) i.e. for each pair 2x60x100=12,000 words. Notice that this eyeballing method is not inconsistent with anything in Kilgarriff's argument: GCY selected 120 contexts in Hansard for each word that DID correspond intuitively to one of the (French) selected senses. It says nothing about any tokens that may have been hard to classify in this way. The figures claimed for the discrimination method against the criterial data vary between 82 and 100% (for different word pairs) of the data for that sense correctly discriminated.

They then move on to a monolingual method that provides sense-tagged data in an unsupervised way. It rests on previous work by Yarowsky (1991) and uses the assignment of a single Roget category (from the 1042) as a sense-discrimination. Yarowsky sense-tagged some of the Grolier corpus in the following way: 100-word contexts for words like "crane" (ambiguous between bird and machinery) are taken and those words are scored by (very roughly, and given interpolation for local context) which of the 1042 Roget categories they appear under as tokens. The sense of a given token of "crane" is determined by which Roget category wins out: e.g. 348 (TOOLS/MACHINERY) for the machinery contexts, one hopes, and category 414 (ANIMALS/INSECTS) for the bird contexts. Yarowsky (1991) claimed 93% correctness for this procedure over a sample of 12 selected words, presumably checked against earlier hand-tagged data.

The interpolation for local effects is in fact very sophisticated and involves training with the 100 word contexts in Grolier of all the words that appear under a given candidate Roget head, a method that they acknowledge introduces some noise, since it adds into the training material Grolier contexts that involve senses of a category 348 word, say, that is not its machinery sense (e.g. crane as a bird). However, this method, they note, does not have the sense-defined-by-language2 problems that come with the

Hansard training method.

In a broad sense, this is an old method, probably the oldest in lexical computation, and was used by Masterman (reported in Wilks 1972) in what was probably the first clear algorithm ever implemented for usage discrimination against Roget categories as sense-criterial. In the very limited computations of those days the hypothesis was deemed conclusive falsified; i.e. the hypothesis that any method overlapping the Roget categories for a word with the Roget categories of neighbouring words would determine an appropriate Roget category for that word in context.

This remains, I suspect, an open question: it may well be that Yarowsky's local interpolation statistics have made the general method viable, and that the 100-word window of context used is far more effective than a sentence. It may be the 12 words that confirm the disambiguation hypothesis at 93% would not be confirmed by 12 more words chosen at random (the early Cambridge work did at least try to Roget-resolve all the words in a sentence). But we can pass over that for now, and head on, to discuss GCY's main claim (2) given the two types of data gathered.

Two very strange things happen at this point as the GCY paper approaches its conclusion: namely, the proof of claim (2) or one-sense-per-discourse. First, the two types of sense-tagged data just gathered, especially the Roget-tagged data, should now be sufficient to test the claim, if a 93% level is deemed adequate for a preliminary test. Strangely, the data derived in the first part of the paper is never used or cited and the reader is not told whether Yarowsky's Roget data confirms or disconfirms (2).

Secondly, the testing of (2) is done purely by human judgement: a "blind" team of the three authors and two colleagues who are confronted by the OALD main senses for one of nine test words, and who then make judgements of pairs of contexts for one of the nine words drawn from a single Grolier article. The subjects are shown to have pretty consistent judgements and, of fifty-four pairs of contexts from the same article, fifty-one shared the same sense and three did not.

Notice here that the display of the OALD senses is pointless, since the subjects are not asked to decide which if any OALD sense the words appear in, and so no Kilgarriff style problems can arise. The test is simply to assign SAME or NOTSAME, and there are some control pairs added to force discrimination in some cases.

What can one say of this ingenious mini-experiment? Lexicographers traditionally distinguish "lumpers" and "splitters" among colleagues: those who tend to break up senses further and those who go for large, homonymic, senses, of which Wierzbicka would be the extreme case. Five GCY colleagues (one had to be dropped to get consistency among the team) from a "lumper" team decided that fifty-one out of fifty-four contexts for a word in a single encyclopaedia article (repeated for eight other words) are in the same sense. Is this significant? I suspect not very, and nothing at all follows to support the myth of discovery that has grown round the paper: the team and data are tiny and not disinterested. The Grolier articles are mini-texts where the hypothesis would, if true, surprise one least. Much more testing is needed before a universal hypothesis about text polysemy enters our beliefs. Of course, they may in the end be right, and all the dogma of the field so far be wrong.

More recently, Yarowsky (1993, 1995) has extended this methodology in two ways: first, he has established a separate claim he calls "one sense per collocation", which is quite independent of local discourse context (which was the separate "one-sense-per-discourse" claim) and could be expressed crudely by saying that it is highly unlikely that the following two sentences (with the "same" collocations for "plants") can both be attested in a corpus:

Plastic plants can fool you if really well made (=organic)
Plastic plants can contaminate whole regions (=factory)

One's first reaction may be to counter-cite examples like "Un golpe bajo" which can mean either a low blow in boxing, or a score one below par, in golf, although "golpe" could plausibly be said to have the same collocates in both cases. One can dismiss such examples (due to Jim Cowie in this case) by claiming both readings are idioms, but that should only focus our mind more on what Yarowsky does mean by collocation.

That work, although statistically impressive, gives no procedure for large-scale sense-tagging taken alone, since one has no immediate access to what cue words would, in general, constitute a collocation sufficient for disambiguation independent of discourse context. An interesting aspect of Yarowsky's paper is that he sought to show that on many definitions of sense and on many definitions of collocation (e.g. noun to the right, next verb to the left etc.) the hypothesis was still true at an interesting level, although better for some definitions of collocation than for others.

In his most recent work (1995) Yarowsky has combined this approach with an assumption that the earlier claim (2: one-sense-per-discourse) is true, so as to set up an iterative bootstrapping algorithm that both extends disambiguating collocational keys (Yarowsky 1993) and retrains against a corpus, while at the same time filtering the result iteratively by assuming (2): i.e. that tokens from the same discourse will have the same sense. The result, on selected pairs (as always) of bi-semous words is between 93 and 97% (for different word pairs again) correct against handcoded samples, which is somewhat better than he obtained with his Roget method (93% in 1991) and better than figures from Schuetze and Pederson (1995) who produce unsupervised clusterings from a corpus that have to be related by hand to intelligible, established, senses. However, although this work has shown increasing sophistication, and has the great advantage, as he puts it, of not requiring costly hand-tagged training sets but instead "thrives on raw, unannotated, monolingual corpora--the more the merrier", it has the defect at present that it requires an extensive iterative computation for each identified bisemous word, so as to cluster its text tokens into two exclusive classes that cover almost all the identified tokens. In that sense it is still some way from a general sense-tagging procedure for full text corpora, especially one that tags with respect to some generally acceptable taxonomy of senses for a word. Paradoxically, Yarowsky was much closer to that last criterion with his 1991 work using Roget that did produce a sense-tagging for selected word pairs that had some "objectivity" predating the experiment.

Although Yarowsky compares his work favorably with that of Schuetze and Pederson in terms of percentages (96.7 to 92.2) of tokens correctly tagged, it is not clear that their lack of grounding for the classes in an established lexicon is that different from Yarowsky, since his sense distinctions in his experiments (e.g. plant as organic or factory) are intuitively fine but pretty ad hoc to the experiment in question and have no real grounding in dictionaries.

Conclusion

It will probably be clear to the reader by now that a crucial problem in assessing this area of work is the fluctuation of the notion of word sense in it, and that is a real problem outside the scope of this paper. For example, sense as between binary oppositions of words is probably not the same as what the Roget categories discriminate, or words in French and English in aligned Hansard sentences have in common.

Another question arises here about the future development of large-scale sense-tagging: Yarowsky contrasts his work with that of efforts like (Cowie et al. 1991) that were dictionary based, as opposed to (unannotated) corpus based like his own. But a difference he does not bring out is that the Cowie et al. work, when optimized with simulated annealing, did go through substantial sentences, mini-texts if you will, and sense-tag all the words in them against LDOCE at about the 80% level. It is not clear that doing that is less useful than procedures like Yarowsky's that achieve higher levels of sense-tagging but only for carefully selected pairs of words, whose sense-distinctions are not clearly dictionary based, and which would require enormous prior computations to set up ad hoc sense oppositions for a useful number of words.

These are still early days, and the techniques now in play have probably not yet been combined or otherwise optimised to give the best results. It may not be necessary yet to oppose, as one now standardly does in MT, large-scale, less accurate, methods, though useful, with other higher-performance methods that cannot be used for practical applications. That the field of sense-tagging is still open to further development follows if one accepts the aim of this paper which is to attack two claims, both of which are widely believed, though not at once: that sense-tagging of corpora cannot be done, and that it has been solved. As many will remember, MT lived with both these, ultimately misleading, claims for many years.

References

Antal, L. (1963) Questions of Meaning. Mouton: The Hague.

Brown, P.F., Di Pietra, S.A., Di Pietra, V.J. and Mercer, R.L. (1991) Word sense disambiguation using statistical methods, Proc. ACL-91.

Bruce, R. and Wiebe, J. (1994) Word-sense disambiguation using decomposable models, Proc. ACL-94.

Copestake, A. and Briscoe, T. (1991) Lexical operations in a unification-based framework, Proc. ACL Siglex Workshop, Berkeley.

Cowie, J., Guthrie, J. and Guthrie, L (1992) Lexical Disambiguation using Simulated Annealing, Proc. Coling-92.

Dagon, I. and Itai, A. (1994) Word sense disambiguation using a second language monolingual corpus, Computational Linguistics, vol. 20.

Gale, W., Church, K. and Yarowsky, D. (1992) One sense per discourse, Proc. 4th DARPA Speech and Natural Language Workshop.

Givon, T. (1967) Transformations of Ellipsis, Sense Development and Rules of Lexical Direction. SP-2896, Systems Development Corp., Sta. Monica, CA.

Green, G. (1989) Pragmatics and Natural Language Understanding. Erlbaum:

Hillsdale, NJ.

Hanks, P. (1994) personal communication.

Ide, N. and Veronis, J. (1994) Have we wasted our time? Proc. International Workshop on the Future of the Dictionary, Grenoble.

Jorgensen, J. (1990) The psychological reality of word senses, Journal of Psycholinguistic Research, vol 19.

Kilgarriff, A. (1993) Dictionary word-sense distinctions: an enquiry into their nature, Computers and the Humanities, vol. 26.

Miller, G. (1985) WordNet: a Dictionary Browser, In Proc. First Internat. Conf. on Information in Data. Waterloo OED Centre, Canada.

Pustejovsky, J. (1995) The Generative Lexicon, MIT Press: Cambridge, MA.

Schuetze, H. and Pederson, J. (1995) Information Retrieval based on Word Sense, Proc. Fourth Annual Symposium on Document Analysis and Information Retrieval. Las Vegas, NV.

Small, S., Cottrell, G., and Tanenhaus, M. (Eds.) (1988) Lexical Ambiguity Resolution, Morgan Kaufmann: San Mateo, CA.

Wierzbicka, A. (1989) Semantics Culture and Cognition, OUP: Oxford.

Wilks, Y. (1972) Grammar, Meaning and the Machine Analysis of Language, Routledge: London.

Wilks, Y. (1978) Making Preferences more Active. Artificial Intelligence, vol. 11.

Wilks, Y., Fass, D., Guo, C.M., McDonald, J., Plate, T., and Slator, B. (1990) Providing machine-tractable dictionary tools. Journal of Machine Translation, vol 5.

Wilks, Y., Slator, B. and Guthrie, L. (1996) Electric Words, MIT Press: Cambridge, MA.

Yarowsky, D. (1991) Word-sense disambiguation using statistical models of Roget's categories, trained on very large corpora, Proc. Coling-92.

Yarowsky, D. (1993) One sense per collocation, Proc. ARPA Human Language Technology Workshop, Princeton.

Yarowsky, D. (1995) Unsupervised word-sense disambiguation rivalling supervised methods, Proc. ACL-95.

Acknowledgement: work referred to was supported by the NSF under grant #IRI 9101232. The paper is also indebted to comments and criticisms from Adam Kilgarriff, David Yarowsky, Karen Sparck Jones, Rebecca Bruce and members of the CRL-New Mexico and University of Sheffield NLP groups. The mistakes are all my own, as always.