Harnessing the Lexicographer in the Quest for Accurate Word Sense Disambiguation

David Tugwell and Adam Kilgarriff

ITRI, University of Brighton, UK

Abstract. This paper outlines a novel architecture for the development of a word sense disambiguation (WSD) system. It is based on the premiss that one way to improve the performance of such systems is through increased, and more flexible, human intervention. To this end a human-WSD program interface, WASPS[1] is being developed for use by lexicographers in organizing corpus data in the drawing up of new dictionary entries. A by-product of this activity will be an accurate sense disambiguation program.

1 Background

Human languages are, to a varying degree, prone to *polysemy*: the phenomenon where the same outward wordform may give rise to different meanings. To illustrate, consider two sentences containing one of the standard examples, the English noun *bank*.

1. He dived into the water and set off strongly towards the opposite *bank*.
2. Requiring traveller's cheques, she popped into the nearest *bank*.

Hearing sentence (1) should be enough to evoke the meaning paraphrasable by *raised strip of land along a river or canal*. Whereas in (2), the meaning *establishment providing financial services* will undoubtedly come to mind. Although going largely unnoticed by human users, this phenomenon places an extra burden on the automatic analysis of language, such as machine translation, information extraction and parsing, so the achievement of high-accuracy sense disambiguation may improve performance in these tasks.[2]

Great efforts have been made over the years to develop automatic techniques to perform this task. An overview of the approaches taken, both statistical and non-statistical, is given in [2]. The recent evaluation exercise SENSEVAL (described in [3]) was a good opportunity to assess current levels of performance and determine future directions. SENSEVAL clearly established that the sense discrimination task in itself was achievable to a high degree of precision by human

[1] A Semi-Automatic Lexicographer's Workbench for Writing Word Sense Profiles, funded by EPSRC.

[2] For a discussion of this point, see [1].

judges (> 95% inter-judge agreement) even where words had a large number of senses.

The best results for automatic systems were achieved by *supervised* systems, ie. those which made use of the varying amounts of pre-tagged training data available in the SENSEVAL task. This data was used to find relevant patterns characteristic of particular senses. However, even with this data, state of the art performance for automatic systems was still significantly lower than human performance, achieving accuracy in the region 75-80%. Furthermore, the performance of such supervised systems appears to cluster in this range, suggesting perhaps that a ceiling on performance may be being reached using existing techniques.

One obvious way to improve disambiguation accuracy is to substantially increase the amount of sense-disambiguated training data available. One obvious problem with this is the prohibitive expense of manually disambiguating training data in the depth and for the range of words required for real-world applications. However, this paper will propose that an alternative course of action might be to harness the skills and energies of another group of people for whom WSD is an important task, namely lexicographers. Modern lexicographers typically draw up dictionary entries with reference to a large on-line database of examples. To make this task less laborious and to reduce the number of examples they have to inspect, they require ways to automatically divide large numbers of example sentences into the senses they are interested in. Given these limitations, the availability of semi-automatic techniques to bring out interesting and typical examples are of great interest to them. This paper will explore how it might be possible to establish a synergy between lexicographers and language engineers serving the interests of both parties and with the aim from the engineering perspective of increasing the potential range and precision of WSD.

2 Method

The central concern of the approach is to make use of human input and decision-making in as efficient way as possible, leaving analysis to automatic techniques whereever possible. The proposed scenario of interaction may be divided into the following stages.

Stage 1: Automatic preanalysis

Given a large corpus[3] containing many instances of the word we are interested in we can find out much about its behaviour using completely automatic techniques. As a first step in disambiguation procedure, we find characteristic patterns which the word occurs in ranked by their statistical significance, that is to what extent the behaviour of this word diverges from that of the vocabulary (or more precisely, the wordclass) as a whole. These patterns include a range of

[3] The WASPS project uses the resource of the 100 million+ word British National Corpus, available with part-of-speech tags.

grammatical relations (found by regular expression matching over the BNC), as well as the looser pattern of *word in proximity*, ie. cooccurrence in an n-word window. The set of patterns is currently under development, but a subset of them is given in the following table, which gives an idea of what partial results of this automatic procedure might look like for the word *bank*.

relation	significant items for *bank*
subject-of	*lend, announce, borrow, refuse...*
object-of	*rob, burst, repay, overflow...*
modifying adjective	*central, opposite, steep, commercial...*
modifying noun	*merchant, river, piggy, canal...*
modifies	*manager, account, holiday, robber...*
preposition *of* + NP head	*England, river, America, Danube*
word in proximity	*money, mud, debt, water...*

It will be seen that these characteristic patterns automatically extracted in this way will in general be relevant to only one of the senses of the word, that is they may be thought of as *clues* for that sense. The assignment of senses to clues is where the human input comes in at the next stage.

Stage 2: Initial interaction with lexicographer

The characteristic patterns (relation-lexeme pairs) for the word in question are then presented to the lexicographer. It is then the lexicographer's task to draw up an initial list of potential senses for the word, which may be modified and extended as the interaction progresses. This could either be done with reference to the automatically-extracted patterns, be based on preexisting dictionaries, or be sensitive to special requirements of the task in hand, for example correspondence to possible targets in the machine translation task.

Having decided on a sense inventory, the lexicographer then runs through the patterns selecting the appropriate sense with the click of a mouse. As will become apparent later this procedure does not have to be lengthy or exhaustive, as even a small number of sense assignments should be sufficient to bootstrap the WSD algorithm described below. To check on the appropriate sense assignment, access is available to the actual sentences in the corpus where the patterns occur.[4]

The patterns with senses assigned are then used as input to the next stage: the sense disambiguation algorithm.

Stage 3: Disambiguation algorithm

The core of the automatic disambiguator is an iterative bootstrapping procedure, based on that described by Yarowsky in [4]. In that paper, Yarowsky uses a

[4] The option is also available to mark individual corpus instances with a particular sense, if so desired, although doing this extensively will slow the process.

few seed collocations, which have been assigned senses, to bootstrap the finding of new good indicators for particular senses. Yarowsky proposes a number of techniques for the initial sense assignment, the one used here is an elaboration of his proposal for limited human intervention to set a small number of seeds. The number to be set here is under the control of the lexicographer in stage 2.

The essential idea of this algorithm is to extract from the seed, "sense-defined", example sentences a decision list of commonly occurring syntactic patterns which are good indicators of a particularly sense. The best clues on the decision list are then used to increase the data set by classifying more of the unclassified examples, and a new decision list is calculated from the increased dataset that results. This process is repeated in a bootstrapping manner until the dataset is divided between the different senses. The resulting decision list can then be applied to disambiguate an unseen dataset. It has certainly been well demonstrated that this is a very effective technique where the number of senses is small in number.

Stage 3a (concurrent with Stage 3): Continued human interaction

In the standard Yarowsky technique the algorithm runs without further human input until all examples are classified into one sense or another. The distinguishing feature of the present approach is the assumption that to achieve a higher level of precision in more difficult conditions (ie. where there are a greater number of senses or less distinct senses to be distinguished) then it will require a more detailed interaction with a lexicographer and possibly a repeated series of interactions to arrive at a high-precision decision list.

To that end new clues for senses are shown to the lexicographer as they are found, so that they can be verified, again with reference to the original sentences from where they derived if necessary. One interesting avenue to explore is that of not finding best clues overall but best clues for each sense and using these, subject to their satisfying some threshold criterion. Clues can be rejected at any time, and the lexicographer is kept informed of the progress of the disambiguation process, that is how many examples have been assigned to each sense and how many remain unassigned.

It is this "rump" of examples that are not easy to assign to any of the senses that may be of great interest to the lexicographer, since it is likely to contain examples of rare senses and uses, of great interest in compiling a dictionary. The task of wading through all examples to find such interesting cases may thus be significantly reduced.

Stage 4: Final outcome

The final outcome of this hopefully relatively brief and non-laborious man-machine interaction may be summarised as follows.

- For the lexicographer, the interaction has been of assistance in preparing the lexical entry: providing a semi-automatic sorting of the examples into

senses, quick and discriminating access to the corpus evidence, and an efficient paring away of standard and repetitive examples.
- For the language engineer, we end up with a decision list of clues for the various senses that can directly be applied to the disambiguation of a new test.
- Finally, we are also left with a potentially reusable resource for other NLP tasks: a (partially) sense-disambiguated corpus.

3 Towards a divide and rule strategy?

One common problem facing word sense disambiguation programs is that the distribution of senses is typically highly skewed, with a few very common senses and a long tail of increasingly rare senses. It is difficult to get sufficient training data to establish good clues for these rarer senses, since they tend to be swamped by the more common ones. However, to the lexicographer these rare senses are equally, if not disproportionately, interesting and important, so it is vital that any automatic techniques that might be employed will bring such examples out of the mass of more common ones. To this end the simple bootstrapping techniques set out above may be modified by aiming for a gradual refinement of senses as the lexicographer interacts with the program.

To give a concrete example of how this might work, let us return to our example word: *bank*. The "financial institution" sense accounts for well over 85% of the $\approx$ 21,000 tokens of *bank* in the BNC, with the sense "bank of waterway" accounting for the majority of the remaining examples. A glance in a dictionary, however, shows a range of other senses such as "ridge of earth" (ie. not by a river), "mass" as in *bank of snow*, "row or series" as in *bank of pumps*, "store" as in *blood bank*, *bank of data*, and even "degree of banking" as in *the plane went into a steep bank*.[5]

One way that we might overcome the swamping effect of overly-dominant senses is to approach the sense division in stages. For example, we could attempt to first make an initial cut of the data between the "financial institution" sense on the one hand and a single grouping of "other senses" on the other. Having classified the patterns in this broad way, removing those examples that can be safely classified as "finance" leaves a smaller set of examples from which statistically-significant patterns may again be calculated and the approach can be applied to this reduced dataset. With the swamping effect of the "financial institution" sense of bank largely removed, there is a good chance that patterns associated with less frequent senses can come to the fore.

[5] There are clearly problems in defining these separate senses with any clarity for the meanings often appear to run into the other without clear boundaries, for example a mound of earth by the side of a river may be thought of as a bank in two different senses, possibly both at the same time. However, although the concept of "word sense" may be undefinable in any absolute way, it may nevertheless still be a useful concept for specific tasks. For discussion of these problems, see for example [5].

It will be seen that the approach provides the flexibility for a number of different approaches, and it remains an empirical problem to discover which are the most effective strategies.

4 Evaluation

As outlined above, the finished system should produce results of various types. In the first instance, it will be possible to judge it for its usefulness as a lexicographic tool, although evaluation of this is bound to have a degree of subjectivity, tests are planned and should provide useful feedback on the design and functionality of the system.

A more objective evaluation will be possible for the resulting disambiguation program, which can be applied to fresh text. Trials have already been carried out with a small subset of the data in the first SENSEVAL competition and preliminary results are encouraging. Starting with the sense inventory provided for the task, and brief human interaction, it seems to be possible to obtain results rivalling the best supervised systems, without using the pre-prepared training data that these systems rely on. The question of whether this performance can be consistently replicated and whether the further addition of "gold standard" training data can increase this performance level awaits further investigation.

5 Conclusion

This paper has presented a proposal for an interactive architecture to improve the accuracy of word sense disambiguation by recruiting the input of lexicographers to the task. Although many of the details of the proposal remain to be worked out, we hope that it can become a flexible tool for all those with an interest in word senses and eventually offer the prospect of increased accuracy in automatic sense disambiguation systems.

References

[1] Kilgarriff, Adam. "What is word sense disambiguation good for?". In Proceedings of NLP Pacific Rim Symposium 1997, Phuket, Thailand.

[2] Ide, Nancy and Jean Véronis. Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics* **24** (1998) 1–40.

[3] Kilgarriff, Adam and Joseph Rosenzweig. English SENSEVAL: report and results. Proceedings of LREC, Athens, May-June 2000 (to appear)

[4] Yarowsky, David. Unsupervised word sense disambiguation rivaling supervised methods. In Proceedings of ACL 33 (1995) 189–196.

[5] Kilgarriff, Adam. "I don't believe in word senses". *Computers and the Humanities* **31 (2)** (1997) 91–113.