Evaluating word sense disambiguation programs: progress report

Adam Kilgarriff
ITRI
University of Brighton

1 Introduction

Evaluation is looming large in the Word Sense Disambiguation (WSD) world. The beneficial impact of a framework for evaluation on parsing, part-of-speech tagging, and information extraction among others is now well documented. Many WSD programs have been written in recent years, and frequently, their authors attempt to evaluate the performance of their system and to compare it to the performance of other systems. However, all would agree that these attempts have left many questions unanswered.

The issues were taken up in a working session on WSD evaluation at the ACL SIGLEX[1] workshop at ANLP 1997 in Washington (Light, 1997). In this paper I report on the working session and likely ways forward for the WSD research community.

Before proceeding, a humble apology: the following discussion, and the workshop itself, mostly proceeded as if English were the only language in the world. To the non-English-speaking world, mea culpa.

2 A Gold Standard corpus

To determine how well a system is performing, a correctly sense-tagged dataset or "gold standard" is required. This involves people sense-tagging a substantial quantity of data (or at least verifying the output of an automatic procedure). Currently, the SEMCOR corpus, developed by the WordNet team,[2] is such a resource, but:

- there are very few word-types for which SEMCOR contains a substantial amount of information. (There are more than 100 sense-tagged corpus instances for just 83 words.)
- WordNet, like any other dictionary, has various shortcomings, and these often result in anomalies in SEMCOR.
- WordNet senses are at a finer-grained level than is appropriate for some NLP tasks.

[1] Special Interest Group on the Lexicon.

[2] (Fellbaum, forthcoming), is to be the primary reference on WordNet, and contains discussions of many of the issues considered at the workshop.

There are also grounds for concern regarding the accuracy of SEMCOR tagging. In general it is hard to obtain high inter-tagger agreement between different human taggers of the same text (Fellbaum et al., 1996). (Ng and Lee, 1996) cite an agreeement rate of 57% between a team of students tagging the SEMCOR data in Singapore, and the official SEMCOR tags.

We may conclude that, in the medium term, WSD needs an improved gold-standard corpus. Unavoidably, this will involve substantial, expensive human tagger effort. A critical question then becomes, how can we use that effort to maximum effect?

3 Resnik and Yarowsky Proposals

Philip Resnik and David Yarowsky presented a paper on WSD evaluation, which served as a starting point for the working session that they led. Here I first summarise their paper, then the discussion in the working session.

Resnik and Yarowsky first make the observations that:

1. WSD evaluation is far from standardised;
2. Different tasks bear different relations to WSD, so, eg., information retrieval may fare best with a quite different approach to WSD to that required for machine translation;
3. Adequately large sense-tagged data sets are hard to obtain;
4. The field has only just begun to narrow down approaches, and identify which ones work well and which do not.

They also explore the contrast between part-of-speech tagging and word sense tagging: for the former, there is a small and fairly standard inventory of tags; within-sentence dependencies are crucial; semantic information is usually not required; and gold standard corpora have long been available. (One might add, high inter-annotator agreement has been attained.) For word senses, none of these hold. In particular, there is no generally-agreed inventory of senses. Different dictionaries carve up the semantic space a word occupies differently. Also, each word has its own sense set, so, where part-of-speech tagging is one problem, sense-tagging a corpus containing 20,000 ambiguous word types is 20,000 problems.[3]

They then made four proposals:

1. **A better evaluation criterion:** Current forays into WSD evaluation mostly allow only exact hits, scoring 1, or anything else, scoring 0. A better scheme would give a positive score to any reduction in the level of ambiguity, so a program which rejected three senses for a five-way ambiguous word, but did not choose between the remaining two (one of which was correct) would get a positive score of less than one.
2. **Minor errors and gross errors:** If *bank* means sand bank in a sentence, then a WSD programme returning the river bank sense is doing better than one returning money-bank. The evaluation metric should reflect this, again assigning a positive score of less than one.

[3] These observations closely parallel those made in (Kilgarriff, 1997).

3. **A framework for common evaluation and test set generation:** This was their detailed proposal about how the community should set about producing a gold standard corpus.

   Each year, a fresh subset of a huge corpus is used; one part of this is reserved for hand-tagging for evaluation, and the remainder, released for training. A sample of, say, 200 ambiguous words (types not tokens) is then chosen to be used for evaluation. Each instance of each of those words in the evaluation subcorpus is manually tagged. The community does not discover what the words are until their software is frozen for evaluation, so there is no risk of the software being optimised for those particular words. A new sample of test-words is selected each year.

   A major concern was that both supervised and unsupervised learning algorithms should be able to use the same evaluation corpus. To this end, any gold-standard, tagged material should be made available as training data to researchers exploring supervised learning methods as soon as this was possible without compromising the "unseen" nature of the evaluation corpus.

4. **A multilingual sense inventory for evaluation:** This is a bid to address the fraught issue of sense inventories. The aim is to apply the principle that if a word has two meanings sufficiently different to receive different translations, then the meanings are treated as distinct senses.

4 Working Session discussion

The working session broadly welcomed the proposals, and all present enjoyed the sense of "we need to get on and do this, as best we can, however imperfect it may be, making whatever compromises we need to make". All were concerned to develop a plan that was workable, both technically and politically, rather than one with theoretical credentials. It would need to command widespread support in the community and to be likely to attract funding. Given the current state of the art in WSD, evaluation will only count as a success if all or most actors approve the method and accept the results.

In the course of the discussion, it became apparent that the central difficulties lay in reaching a consensus between two cultures – the computer scientists, who view a set of dictionary definitions as data they are to work with (and would like to be able to treat them as fixed) and the humanists, who had detailed experience of lexicography, textual analysis and similar, and whose dominant concern lay in the sheer difficulty of identifying and defining word senses.

The thesis came from Resnik and Yarowsky, in the computer scientists' camp. The antithesis was that:

- it was hard to get high inter-rater agreement. Without that, the gold standard would be fool's gold;

- it was impossible for a human to tag consistently and coherently **unless** the dictionary they were using was

  - well written,
  - made sense distinctions intelligently and clearly,

  - provided examples –preferably several– of the word being used in each sense,

  and had well-defined policies on, *inter alia*,

  - the level of granularity of each sense;
  - collocations and multi-word expressions;
  - nesting of senses;
  - regular polysemy;
  - partially-conventionalised metaphor and metonymy.

- even with the an ideal dictionary, there will be uses of a word for which no dictionary sense fits, or more than one does.

As Bob Amsler pointed out, the sense-tagging task is the dual of what the lexicographer does. The lexicographer takes corpus instances of a word and puts them into separate heaps, calls each heap a distinct word sense, and writes a definition for it. The sense-tagger is given the definitions for each heap and allocates corpus instances to them. The validity of each task is constrained by the validity of the other.

4.1 Dialectic in action (1): Multiple correct answers

> Should a human tagger be allowed to say that more than one sense of a word applies to a corpus instance of the word (so there are multiple correct answers in the gold standard corpus)?

The computer scientists were initially unenthusiastic, since it makes the gold standard harder to use, and performance statistics more complex to define and interpret. But the humanists were adamant that sometimes, multiple correct answers were simply the truth of the matter, and the cost of defining this possibility away was that the gold standard would not be true. The computer scientists then started considering more sophisticated evaluation measures, which could provide scoring schemes for multiple correct answers. (There was some discussion of the relation of this question to the question of grain-size, since the issue arises where a more specific and a more general sense are both valid for a corpus instance). The matter went to the vote and it was agreed that multiple correct answers should be retained as a possibility, though the human taggers should be discouraged from giving multiple answers unless they were clear that a single answer would be untrue.

4.2 Dialectic in action (2): Tagger-lexicographers

If the computer scientists' bottom line was that scoring had to be possible, the humanists' was that the lexicography must not be immutable. Where the taggers found that it wass impossible to tag the corpus lines for a word accurately, because the dictionary entry they were working to was wrong, or vague, or incomprehensible, then it had to be possible for them to force a revision to the dictionary entry.[4]

[4] In the course of producing SEMCOR, the WordNet team had found it necessary to introduce such a feedback loop. The paradigm of simultaneous lexicography and sense-tagging has been familiar for a while to lexicographers (Atkins, 1993).

Resnik and Yarowsky's thesis already contained the proposal that manual tagging and evaluation should be performed on a limited set of word types each year (see also (Kilgarriff, 1997; Kilgarriff and Scott, 1997)), and this made it viable to integrate the humanists' demands for revisable sense-distinctions into the overall plan. If this set of word types were to be held at 200 each year, then the brief for the team of human taggers would be to amend the dictionary entry where the corpus evidence showed the existing entry to be unworkable, as well as to produce the sense-tagged corpus. The team of taggers would become a team of tagger-lexicographers.

A convergent development is the WordNet team's KILO project in which the WordNet entries for a thousand (hence KILO) words are examined and improved, using corpus-based methods, with many examples going into the lexicon for each sense. Some of the words used for WSD evaluation could be taken from the KILO thousand, and for them, the 'extended lexicography' and sense-tagging task would already be underway.

The mutability of the dictionary would have repercussions for several aspects of the Resnik-Yarowsky proposals. First, the final state of the dictionary would not be fixed until a short while before the evaluation, which would create difficulties for those training strategies which depend on substantial pre-compilation of the dictionary. Secondly, the segment of the lexicon to be used for the evaluation will be systematically different from the remainder of the lexicon: it will have more extensive entries, with more examples and more nesting. Thirdly, the original proposal argued that the test-set of word-types should be unseen, but it is unlikely that this could be kept to, since, as soon as the 'frozen' dictionary was made public, any user could establish that those words with changed definitions were likely to be in the test set. My interpretation of the tenor of the meeting was that it accepted the need for these compromises.

5 Recommendations

The workshop was a very useful consensus-building exercise. The basic Resnik-Yarowsky strategy (as in their proposal 4) was agreed upon, with the following qualifications and further specifications.

Improved metric As covered in Resnik-Yarowsky proposals 1 and 2, but further developed to take account of multiple correct answers.

Tagger-lexicographers Lexicon to be developed, as discussed above.

Part of speech tagging The WSD task should be evaluated in isolation from the part-of-speech (POS) tagging task. This would mean that researchers working on WSD did not also need to apply themselves to POS-tagging, and also that good solutions for WSD would not score badly owing to low-quality POS. The manual taggers would ensure that the gold standard corpus was correctly POS-tagged as well as sense-tagged, and for evaluation purposes, POS of the word to be disambiguated would be available as part of the input (which programs would of course be free to ignore if they wished to do their own POS-tagging).

Dictionary for sense inventory WordNet has the great merits, from the research community's perspective, of being free, without licensing constraints, and available by `ftp`. WordNet versions for numerous other languages are currently under development and will be available before very long, in most cases still free and without restrictions. The WordNet team is enthusiastic about collaborating with the WSD and NLP community in its further development. WordNet is now the most widely used lexical resource. US Government funding has gone into WordNet, so the US Government is likely to look favourably on it being used for other purposes of which it approves. For these reasons, it was decided that WordNet should be the starting point for the sense inventory to be used for the manual tagging.

Grain size As Yorick Wilks pointed out, both researchers and funders will be happier if good programs score 90%, than if they score 30%, and we would do well to bear this in mind in defining the scoring metric! The state of the art is such that WSD with coarse-grained senses, or homographs, is attainable at well over 90%, whereas for fine-grained WSD even human inter-annotator agreement is very low. Thus the 'headline' figure that the scoring metric defines should be a score for coarse-grained disambiguation.

Different NLP applications require WSD for different purposes, and coarse-grained disambiguation may well be sufficient for Information Retrieval or to resolve parsing ambiguities, but inadequate for Machine Translation. Thus, in addition to the headline score, the metric should provide a measure for success at finer-grained disambiguation.

This imposes a constraint on the dictionary and the tagging. The dictionary must use nested senses, so that "coarse-grained" and "fine-grained" are defined, and the taggers will require guidelines on when to assign a fine-grained tag to a corpus instance, when a coarse-grained tag, and when both. There is a substantial further lexicographic task here, as currently, there is very little nesting of senses in WordNet. (Whereas the traditional dictionary entry lends itself to representing nested senses, the basic organisational mechanism in WordNet, the synset, does not.) Sense nesting is currently being introduced in WordNet, and the next release will contain some nesting information for many nouns and verbs, but nonetheless, clustering WordNet's fine-grained senses into coarse senses will in all likelihood be a substantial task for the tagger-lexicographers.

General or Special language While there were arguments for using a sublanguage corpus for evaluation – the problem is simpler and better defined, and probably the bulk of applications using WSD in the medium term future will be sub-language-based – it was agreed that a general language corpus should be used. This would avoid the possible 'ghettoisation' of the evaluation exercise and ensure the maximum generality of the algorithms. For some potential WSD clients, such as web browsers, the ability to deal with general language was essential. It was noted that WSD strategies geared around tuning a lexicon for a specific sublanguage (eg., (Basili, Della Rocca, and Pazienza, 1997)) would not be readily integrated into a general-language evaluation scheme.

6 Next move

There was a discussion as to whether the field was ready to bid for a MUC-style competitive evaluation: the outcome was that it was not. The PARSEVAL model, in which interested parties met to define the evaluation scheme, which was developed alongside attempts at using it, was preferred, at least for the time being.

The WSD evaluation exercise is unlikely to be able to proceed without grant funding to employ the tagger-lexicographers. Questions of whether the tagger-lexicographers might all be in one place, or distributed, and how the grant proposal might be put together, were only briefly touched on.

It was agreed that SIGLEX was the body to take the matter forward. Further postings and discussions will take place on the SIGLEX mailing list, with another meeting in one or two years.

Disclaimer

The interpretation of discussions at the workshop is my own. While I have done my best to represent views and decisions accurately, there has not been sufficient time to check with all those present. The organisers' report of the working session will be available shortly.

References

Atkins, Sue. 1993. Tools for computer-aided lexicography: the Hector project. In *Papers in Computational Lexicography: COMPLEX '93*, Budapest.

Basili, Roberto, Michelangelo Della Rocca, and Maria Teresa Pazienza. 1997. Towards a bootstrapping framework for corpus semantic tagging. In Light, 1997, pages 66–73.

Fellbaum, Christiane, editor. Forthcoming. *WordNet: An Electronic lexical Database.* MIT Press.

Fellbaum, Christiane, Joachim Grabowski, Shari Landes, and Andrea Baumann. 1996. Matching words to senses in WordNet: Naive *vs.* expert differentiation of senses. In Fellbaum, forthcoming.

Kilgarriff, Adam. 1997. Sample the lexicon. Technical Report ITRI-97-01, ITRI, University of Brighton.

Kilgarriff, Adam and Donia Scott. 1997. Word sense profiles. Grant application to the UK Engineering and Physical Sciences Research Council.

Light, Marc, editor. 1997. *Tagging Text with Lexical Semantics: Why, What and How?*, Washington, April. SIGLEX (Lexicon Special Interest Group) of the ACL.

Ng, Hwee Tou and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *ACL Proceedings*, June.

Resnik, Philip and David Yarowsky. 1997. A perspective on word sense disambiguation methods and their evaluation. In Light, 1997.