95% Replicability for Manual Word Sense Tagging

Adam Kilgarriff
ITRI, University of Brighton, Lewes Road, Brighton UK
email: adam@itri.bton.ac.uk

People have been writing programs for automatic Word Sense Disambiguation (WSD) for forty years now, yet the validity of the task has remained in doubt. At a first pass, the task is simply defined: a word like *bank* can mean 'river bank' or 'money bank' and the task is to determine which of these applies in a context in which the word *bank* appears. The problems arise because most sense distinctions are not as clear as the distinction between 'river bank' and 'money bank', so it is not always straightforward for a person to say what the correct answer is. Thus we do not always know what it would mean to say that a computer program got the right answer. The issue is discussed in detail by (Gale et al., 1992) who identify the problem as one of identifying the 'upper bound' for the performance of a WSD program. If people can only agree on the correct answer x% of the time, a claim that a program achieves more than x% accuracy is hard to interpret, and x% is the upper bound for what the program can (meaningfully) achieve.

There have been some discussions as to what this upper bound might be. Gale et al. review a psycholinguistic study (Jorgensen, 1990) in which the level of agreement averaged 68%. But an **upper** bound of 68% is disastrous for the enterprise, since it implies that the best a program could possibly do is still not remotely good enough for any practical purpose.

Even worse news comes from (Ng and Lee, 1996), who re-tagged parts of the manually tagged SEMCOR corpus (Fellbaum, 1998). The taggings matched only 57% of the time.

If these represent as high a level of inter-tagger agreement as one could ever expect, WSD is a doomed enterprise. However, neither study set out to identify an upper bound for WSD and it is far from ideal to use their results in this way. In this paper we report on a study which did aim specifically at achieving as high a level of replicability as possible.

The study took place within the context of SENSEVAL, an evaluation exercise for WSD programs.[1] It was, clearly, critical to the validity of SENSEVAL as a whole to establish the integrity of the 'gold standard' corpus against which WSD programs would be judged.

Measures taken to maximise the agreement level were:

- humans: whereas other tagging exercises had mostly used students, SENSEVAL used professional lexicographers
- dictionary: the dictionary that provided the sense inventory had lengthy entries, with substantial numbers of examples
- task definition: in cases where none, or more than one, of the senses applied, the lexicographer was encouraged to tag the instance as "unassignable" or with multiple tags[2]

[1] The exercise is chronicled at `http://www.itri.bton.ac.uk/events/senseval` and in (Kilgarriff and Palmer, Forthcoming), where a fuller account of all matters covered in the poster can be found.

[2] The scoring algorithm simply treated "unassignable" as another tag. (Less than 1% of instnaces were tagged "unassignable".) Where there were multiple tags and a partial match between taggings, partial credit was assigned.

- arbitration: first, two or three lexicographers provided taggings. Then, any instances where these taggings were not identical were forwarded to a third lexicographer for arbitration.

The data for SENSEVAL comprised around 200 corpus instances for each of 35 words, making a total of 8455 instances. A scoring scheme was developed which assigned partial credit where more than one sense had been assigned to an instance. This was developed primarily for scoring the WSD systems, but was also used for scoring the lexicographers' taggings.

At the time of the SENSEVAL workshop, the tagging procedure (including arbitration) had been undertaken once for each corpus instance. We scored lexicographers' initial pre-arbitration results against the post-arbitration results. The scores ranged between 88% to 100%, with just five out of 122 results for <lexicographer, word> pairs falling below 95%.

To determine the replicability of the whole process in a thoroughgoing way, we repeated it for a sample of four of the words. The words were selected to reflect the spread of difficulty: we took the word which had given rise to the lowest inter-tagger agreement in the previous round, (*generous*, 6 senses), the word that had given rise to the highest, (*sack*, 12 senses), and two words from the middle of the range (*onion*, 5, and *shake*, 36). The 1057 corpus instances for the four words were tagged by two lexicographers who had not seen the data before; the non-identical taggings were forwarded to a third for arbitration. These taggings were then compared with the ones produced previously.

The table shows, for each word, the number of corpus instances (Inst), the number of multiply-tagged instances in each of the two sets of taggings (A and B), and the level of agreement between the two sets (Agr).

There were 240 partial mismatches, with partial credit assigned, in contrast to just 7 complete mismatches.

A instance on which the taggings disagreed was:

Give plants *generous* root space.

Word	Inst	A	B	Agr %
generous	227	76	68	88.7
onion	214	10	11	98.9
sack	260	0	3	99.4
shake	356	35	49	95.1
ALL	1057	121	131	95.5

Sense 4 of *generous* is defined as simply "abundant; copious", and sense 5 as "(of a room or building) large in size; spacious". One tagging selected each. In general, taggings failed to match where the definitions were vague and overlapping, and where, as in sense 5, some part of a defintion matches a corpus instance well ("spacious") but another part does not ("of a room or building").

Conclusion

The upper bound for WSD is around 95%, and Gale et al.'s worries about the integrity of the task can be laid to rest. In order for manually tagged test corpora to achieve 95% replicability, it is critical to take care over the task definition, to employ suitably qualified individuals, and to double-tag and include an arbitration phase.

References

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Mass.

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.

Julia C. Jorgensen. 1990. The psychological reality of word senses. *Journal of Psycholinguistic Research*, 19(3):167–190.

Adam Kilgarriff and Martha Palmer. Forthcoming. Guest editors, Special Issue on SENSEVAL: Evaluating Word Sense Disambiguation Programs. *Computers and the Humanities*.

Hwee Tou Ng and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *ACL Proceedings*, pages 40–47, Technical University, Berlin, Santa Cruz, California.