No-bureaucracy evaluation

Adam Kilgarriff
ITRI, University of Brighton
adam@itri.brighton.ac.uk

SENSEVAL is a series of evaluation exercises for Word Sense Disambiguation. The core design is in accordance with the MUC and TREC model of quantitative, developer-oriented (rather than user-oriented) evaluation. The first was in 1998, with tasks for three languages and 25 participating research teams, the second in 2001, with tasks for twelve languages, thirty-five participating research teams and over 90 participating systems. The third is currently in planning. The scale of the resources developed is indicated in Table 1 (reproduced from (Edmonds and Kilgarriff, 2002)).[1]

In this paper we address five of the workshop themes from a SENSEVAL perspective:

1. organisational structure
2. re-use of corpus resources: pro and con
3. the web and evaluation
4. SENSEVAL and Machine Translation evaluation
5. re-use of metrics: a cautionary tale.

1 Organisation

One aspect of SENSEVAL of interest here is its organizational structure. It has no centralised sponsor to fund or supply infrastructure. Almost all work was done by volunteer effort with just modest local grant funding for particular subtasks, with organisers answerable to no-one beyond the community of WSD researchers. This was possible because of the level of commitment. People wanted the evaluation framework, so they were willing to find the time, from whatever slack they were able to concoct.

At the SENSEVAL-1 workshop, the possibility of finding an official sponsor –most likely the EU or a branch of the US administration– was discussed at length and vigorously. The prevailing view was that, while it was nice to have more money around, it was not necessary and came at a cost. Various experiences were cited where researchers felt their energies had been diverted from the research itself to the processes of grant applications, cost statements, and the strange business of writing reports which in all likelihood no-one will ever read. My experience, as co-ordinator of SENSEVAL-1 and chair of SENSEVAL-2, was that, without external funding but with great goodwill and energy for the task at various locations round the globe, it was possible to get a vast amount done in a short time, at some cost to family life but with a minimum of misdirected effort.

At several points, potential funders have said "All you need to do is fill in our form..." It is always worth asking whether this is a poisoned chalice. How much effort will it take to fill in, and how much more to follow it through? What is the cost to my engagement and enthusiasm of doing things their way (as I shall have to, if I take the king's shilling, as good governance demands that procedures are followed, forms are filled, any changes to the original plan are justified and documented ...).

I should note that, possibly, my perspective here is atypical. As the co-ordinator, without

[1] SENSEVAL data sets and results are available at http://www.senseval.org

Table 1: SENSEVAL-2, resources, participation, results.

Language	Task[a]	Systems	Lemmas	Instances[b]	IAA[c]	Baseline[d]	Best score
Czech	AW	1	–[e]	277,986	–	–	94
Basque	LS	3	40	5,284	75	65	76
Dutch	AW	1	1,168	16,686	–	75	84
English	AW	21	1,082	2,473	75	57	69
English	LS	26	73	12,939	86	48/16[f]	64/40
Estonian	AW	2	4,608	11,504	72	85	67
Italian	LS	2	83	3,900	21	–	39
Japanese	LS	7	100	10,000	86	72	78
Japanese	TM	9	40	1,200	81	37	79
Korean	LS	2	11	1,733	–	71	74
Spanish	LS	12	39	6,705	64	48	65
Swedish	LS	8	40	10,241	95	–	70

[a] AW: all-words task, LS: lexical sample, TM: translation memory.

[b] Total instances annotated in both training and test corpora. In the default case, they were split 2:1 between training and test sets.

[c] Inter-annotator agreement is generally the average percentage of cases where two (or more) annotators agree, before adjudication. However there are various ways in which it can be calculated, so the figures in the table are not all directly comparable.

[d] Generally, choosing the corpus-attested most frequent sense, although this was not always possible or straightforward.

[e] A dash '–' indicates the data was unavailable.

[f] Supervised and unsupervised scores are separated by a slash.

a funder as taskmaster, I had a particularly free hand to ordain as I saw fit. This was most agreeable, but it is quite possible that others involved saw me as their (more or less reasonable, more or less benevolent) dictator and bureaucracy, and did not share the pleasures of autonomy that I experienced.

I am not sure that I advocate the no-bureaucracy approach: clearly, it depends on there being some slack somewhere which can be redirected. It is however a model well worth considering, if only because it is such fun working with other committed volunteers for no better reason than that you all want to reach the same goal.

2 The re-use trap

Consider the following position (Redux, 2001):

> As followers of the literature will have noted, great strides have been made in statistical parsing. In two decades, system performance figures have soared to over 90%. This is a magnificent tale. Parsing is cracked. An enormous debt is owed to the producers of the Penn Treebank. As anticipated by Don Walker, marked-up resources were what we needed. Once we had them, the algorithm boys could set to work, and whoomph!
>
> The benefits of concentrating on the one corpus have been enormous. The field has focused. It has been the microscope under which the true nature of language has become apparent. Like Mendel unpacking the secrets of all species' genetics through assiduous attention to sweet peas, and sweet peas alone, Charniak, Collins, and others have unpacked the secrets of grammatical structure through rigorous attention to the Wall Street Journal.
>
> We would now like to point out the unhelpfulness of comments appear-

> ing on the CORPORA mailing list, reporting low performance of various statistical POS-taggers when applied to text of different types to the training material, and also of a footnote to a recent ACL paper, according to which a leading Penn-Treebank-trained parser was applied to literary texts but then its performance "significantly degraded". These results have not, I am glad to say, entered beyond that footnote into the scientific literature. The authors should realise that it is *prima facie* invalid to apply a resource trained on one type of data, to another. Anyone wishing to use a statistical parser on a text type for which a manually-parsed training corpus does not exist, must first create the training corpus. If they are not willing to do that, they may as well accept that ten years of dazzling progress is of no use to them.
>
> ...
>
> So now, our proposal. We are encouraged to see the amount of work based on the Wall Street Journal which appears in ACL proceedings. However we remain concerned about the quantity of papers appearing there which fail to use a rigorous methodology, and fail to build on the progress outlined above. These papers tend to fall outside the domain which has become the testing ground for our understanding of the phenomenon of language, viz, the Wall Street Journal. Outside the Wall Street Journal, we are benighted. May I suggest that ACL adopt a policy of accepting only papers investigating the language of the Wall Street Journal.

A similar position was discussed in relation to SENSEVAL. There was a move to use, in part or in whole, the same sample of words (ca 40 items) for SENSEVAL-2 (English lexical sample task) as had been used in SENSEVAL-1. This would have promoted comparability of results across the two exercises. However, we were anxious about continuing to focus our efforts on just 40 of the 10,000 ambiguous words of the language, as it seemed plausible that some issues had simply not arisen in the first sample, and if we did not switch sample, there was no chance that they would ever be encountered.

All SENSEVAL resources are in the public domain and can be (and have been) used by researchers wanting to compare their system performance with performance figures as in SENSEVAL proceedings. Of course such comparison will never be fair, as systems competing under the examination conditions of the evaluation exercise were operating under time pressure, and did not always have time to correct even the most egregious of bugs. However it is hard to see how the evaluation series can keep the sheer range and variety of language use on the agenda if samples are reused.

3 Language flow and the web

> You cannot step twice into the same river, for other waters are constantly flowing on.
> *Heraclitus* (c. 535-c. 475 BC)

We are currently planning a SENSEVAL-3 task where the test data will be instances of words in web pages, as located by a search engine. Test data will be defined by URL, line number and byte offset. The goal is to explore what happens when laboratory conditions are changed for web conditions. It will support exploration of how supervised-training systems perform when test set and training set are no longer subsets of the same whole. Partipants will be expected to first retrieve the web page and then apply WSD to it. This will allow systems to use a wider context than is possible in the usual paradigm of short-context test instances. They could, for example, gather a corpus of the reference URL, plus any pages it links to, plus other pages close to it in its directory tree, in order to identify the domain of the instance. In general, it makes space

for a range of techniques which the SENSEVAL paradigm to date has ruled out.

Clearly, web pages may change or die between selecting URLs for manual tagging at set-up time, and the evaluation period, resulting in wasted manual-tagging effort. We shall minimize the waste by, first, drawing up a candidate list of URL's, then, checking them to see whether they are still available and unchanged a month or so later. The fact that some web pages have died will not invalidate the exercise. It just means there will be fewer usable test instances than test-URLs distributed.

One hypothesis to be explored is that supervised-training systems are less resilient than other system-types, in the real world situation where the data to be disambiguated "in anger" may not match the text type of the training corpus. The relation between the performance of supervised-training systems in the laboratory and in the wild is to my mind one of the critical issues at the current point in time, given the ascendancy that the paradigm has achieved in CL.

It may also shed light on the relation between a linguistic/collocational view of word senses and one dominated by domain. Inevitably, for some words, there will be a poor match between the domains of training-corpus instances and the domains of web instances. While this might seem 'unfair' and a problem following from the biases of the web, it is a fact of linguistic life. The concept of an unbiased corpus has no theoretical credentials. The task will explore the implications of working with a corpus whose biases are unknown, and in any case forever changing.

The web also happens to be the corpus that many potential customers for WSD need to operate on, so the task will provide a picture of whether WSD technology is yet ready for these potential clients.

4 SENSEVAL and Machine Translation evaluation

As noted above, overall SENSEVAL design is taken from MUC. We have also followed MUC and TREC discussions of the hub-and-spokes model and the need to forever look towards updating the task, to guard against participants becoming expert at the task as defined but not at anything else.

WSD is not a task of interest in itself. One does WSD in order to improve performance on some other task. The critical end-to-end task, for WSD, is Machine Translation (Kilgarriff, 1997).

In SENSEVAL-2, for Japanese there was a translation memory task, which took the form of an MT evaluation (Kurohashi, 2001). In that experimental design, each system response potentially requires individual attention from a human assessor. As in assessing human or computer translation, one cannot specify a complete set of correct answers ahead of time, so one must be open to the possibility that the system response is correct but different from all the responses seen to date. Thus the exercise is potentially far more expensive than the MUC model. In the MUC model, human attention is required for each data instance. In this model, human attention is potentially required for each data-instance/system combination.

Another consequence is that there is no free-standing, system-independent gold standard corpus of correct answers. New or revised systems cannot simply test against a gold standard (unless they limit their range of possible answers to ones already encountered, which would introduce further biases).

So it is a more complex and costly form of evaluation. However it is also far more closely related to a real task. It is a direction that SENSEVAL needs to take.[2] The MUC-style fixed-sense-inventory should be seen as what was necessary to open the chapter on WSD evaluation: a graspable, manageable task when we had no experience of the difficulties we might encounter, which also provided researchers with some objective datasets for their development work. For the future the

[2]It is also the route we have taken in the WASPS project, which is geared towards WSD for MT (Koeling et al., 2003).

emphasis needs to be on assessments such as the Japanese one, related to real tasks.

5 Metric re-use: kappa

Consider the (fictional) game show "Couples". The idea is to establish which couples share the same world view to the greatest extent. Each member of the couple is put in a space where they cannot hear what the other is saying, and is then asked twenty multiple-choice questions like

What is the greatest UK pop group of the 1960s?

The Beatles/The Rolling Stones

or

Which month is your oldest nephew/niece's birthday?

Jan/Feb/Mar/Apr/May/Jun/Jul /Aug/Sep/Oct/Nov/Dec /No-nephew-or-niece

The couple that gives the same answer most often wins.

Different couples get different questions, sometimes with different numbers of multiple-choice options, and this introduces a risk of unfairness. If one couple gets all two-way choices, while another gets all 13-way choices, and both agree half the time, the 13-way couple have really done much better. Random guessing would have got (on average) a 50% score for the couple who got the two-way questions, whereas it would only have got a 1/13 or 7.7% score for the others.

One way to fix the problem is to give, for each question, not a full point but a score modified to allow for what random guessing would have given. This can be defined as

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

where $P(A)$ is the proportion of times they actually agree, and $P(E)$ is the proportion of times they would agree by chance.

This is called the Kappa statistic. It was developed within the discipline of Content Analysis, and introduced into the HLT world by Jean Carletta (Carletta, 1996).

Inter-Annotator Agreement

For HLT, the issue arises in manual tagging tasks, such as manually identifying the word class or word sense of a word in the text, or the discourse function of a clause. In each of these cases, there will be a fixed set of possible answers. Consider two exercises, one where a team of two human taggers tag a set of clauses for discourse function using a set of four possible functions, the other where another team of two uses a set of fifteen possible functions. If the first team gave the same answers 77% of the time, and the second gave the same answers 71% of the time, then, at a first pass, the first team had a higher agreement level. However they were using a smaller tagset, and we can use kappa to compensate for that. The kappa figure for the first team is

$$\frac{0.77 - 1/4}{1 - 1/4} = \frac{0.52}{0.75} = 0.69$$

and that for the second team is

$$\frac{0.71 - 1/15}{1 - 1/15} = \frac{0.64}{0.93} = 0.69$$

The inter-annotator agreement (IAA) can be presented as simple agreement figures of 77% and 71%, or as kappa values of 0.69 in both cases.

IAA matters to HLT evaluation because human tagging is what is needed to produced 'gold standard' datasets against which system performance can be judged. The simplest approach is for a person to mark up a text, and to evaluate the system against those taggings. But the person might make mistakes, and there may be problems of interpretation and judgement calls where a different human may well have given a different answer. So, for gold standard dataset development, each item to be tagged should be tagged by at least two people.

How confident can we be in the integrity of the gold standard? Do we really know that it is correct? A central consideration is IAA: if taggers agreed with each other nearly all the time, we can be confident that, firstly, the gold

standard corpus is not full of errors, and secondly, that the system of categories, or tags, according to which the markup took place is adequate to the task. If the tags are not well-suited to the task and adequately defined, it will frequently be arbitrary which tag a tagger selects, and this will show up in low IAA.

Reservations

Carletta presented kappa as a better measure of IAA than uncorrected agreement. In the specific cases she describes, this is certainly valid.

Those cases are very specific. Kappa is relevant where the concern is that an IAA figure based on a small tagset is being compared with one based on a large tagset. Where that is the focus of the investigation, kappa is an appropriate statistic.

Where it is not, there are arguments for and against the use of kappa. In its favour is that it builds in compensation for distortions that might otherwise go unnoticed resulting from different tagset sizes.

Against is, principally, the argument that kappa figures are hard to interpret. A simple agreement figure is just that: it is clear what it means, and the critical question of whether, say, 90% agreement is 'good enough' is one for the reader to form their own judgment on. With a kappa figure of .85, the reader needs to, firstly, understand the mathematics of kappa, and secondly, bear in mind the various complexities of how kappa might have been calculated (see also below), before forming a judgment. To "help" the reader with this task, there are various discussions in the literature as to how different kappa figures are to be interpreted. Sadly, these are contradictory (and even if they weren't, it is the duty of any critical reader to form their own judgment on what is good enough.)

Complexities in the calculation

Above we present kappa in its simplest form. Naturally, when used in earnest additional issues arise. The observations below arose principally from the consideration of how we might use kappa in SENSEVAL. The task was to produce a gold standard corpus in which words were associated with their appropriate meanings, with the inventory of meanings taken from a dictionary.

Firstly, tagset size is assumed to be fixed. In the SENSEVAL context, there were three issues here.

1. There were two variants of the task: 'lexical sample' and 'all-words'. In the all-words variant, all content words in a text are tagged. Some will be highly polysemous, others not polysemous at all. It is not clear how to present kappa figures that are averages across datasets where the tagset size varies.

   In the lexical sample task, first, a sample of sentences containing a particular word is identified, and then, only the instances of that word are tagged, so the issue does not arise immediately. It does still arise if a kappa figure is to be computed which draws together data from more than one lexical-sample word.

2. In addition to the dictionary senses for the word, there were two tags, U for 'unassignable' and P for 'proper name', which were always available as options for the human taggers. If included, for purposes of calculating kappa, a word that only has two dictionary senses is classified as a four-way choice, which seems inappropriate, particularly as U and P tags were quite rare and absent entirely for some words.

3. There were a number of other 'marginal' senses which, if included in the tag count, extend it greatly (for some words). In the SENSEVAL-1, taggers largely worked within a given word class, so noun instances of *float* were treated separately from verb instances, but, in e.g., noun cases where none of the noun instances fitted, they were instructed to consider whether any of the verb senses were a good semantic match (even though they

evidently could not be a syntactic match). Also some words formed part of numerous multi-word units that were listed in the dictionary. Where a tagger found the lexical-sample word occurring within a listed multi-word unit, the instruction was to assign that as a sense.

One response to issues 2 and 3 is to use a more sophisticated model of random guessing, in which, rather than assuming all tags are equally likely for the random guesser, we use the relative frequencies of the different tags as the basis for a probability model . The method succeeds in giving less weight to marginal tags, at the cost of making the maths of the caluclation more complex and the output kappa figures correspondingly harder to interpret.

Secondly, the SENSEVAL tagging scheme allowed human taggers to give multiple answers, and also allowed multiple answers in the tagging scheme.

Thirdly, in SENSEVAL the number of humans tagging an instance varied (according to whether or not the instance was problematic).

Fourthly, there is a distinction between two kinds of occasion on which two taggers give different tags. It may be a problematic case to tag, or it may be simple human error (such as a typo). Arguably, simple typos and similar are of no theoretical interest and should be corrected before considering IAA. A related point is the distinction between agreement levels (between individual taggers) and replicability (between teams of taggers). Where the concern is the integrity of a gold standard resource, replicability is the real matter of interest: would another team of taggers, using the same data, guidelines and methods, arrive at the same taggings? A tagging methodology which guards against simple errors, wayward individuals, and wayward interpretations will tend to produce replicable datasets.

All of these considerations can be addressed using a variant of kappa. My point is that kappa becomes harder and harder to interpret, as more and more assumptions and intricacies are built into its calculation.

Kappa has been widely embraced as an example of an aspect of evaluation technology that carries across different HLT evaluation tasks, giving a shimmer of statistical sophistication wherever it alights. My sense is that it is a bandwagon, which HLT researchers have felt they ought to jump on in order to display their scientific credentials and ability to use statistics, which, in many places where it has been used, has led to little but gratuitous obfuscation.

6 Conclusion

Clearly, we would like new HLT evaluation exercises to benefit from evaluation work already done. This paper explores several issues that have arisen from the SENSEVAL experience.

References

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Philip Edmonds and Adam Kilgarriff. 2002. Introduction to the special issue on evaluating word sense disambiguation systems. *Journal of Natural Language Engineering*, 8(4).

Adam Kilgarriff. 1997. What is word sense disambiguation good for? In *Proc. Natural Language Processing in the Pacific Rim (NLPRS '97)*, pages 209–214, Phuket, Thailand, December.

Rob Koeling, Roger Evans, Adam Kilgarriff, and David Tugwell. 2003. An evaluation pf a lexicographer's workbench: building lexicons for machine translation. In *EACL workshop on resources for Machine Translation*, Budapest.

Sadao Kurohashi. 2001. SENSEVAL-2 japanese translation task. In *Proc.* SENSEVAL-*2: Second International Workshop on Evaluating WSD Systems*, pages 37–40, Toulouse, July. ACL.

Swift Redux. 2001. A modest proposal. *ELSnews*, 10(2):7.