Designing a task for SENSEVAL-2

Philip Edmonds

May 16, 2000

WHAT YOU NEED TO DO

This document provides guidelines for setting up and running a task for SENSEVAL-2. These guidelines are based on experience gathered from SENSEVAL-1, which involved individually organized tasks in English (Kilgarriff and Rosenzweig 2000), French (Segond 2000), and Italian (Calzolari and Corazzari 2000), and on the evaluation protocol proposed by Resnik and Yarowsky (1999). Although Resnik and Yarowsky's proposal assumes that a very large corpus is available and that the same type of evaluation is run repeatedly over several years, we suggest that a variation of it be adopted, as outlined below. As these are merely guidelines, if your circumstances require you to adopt a different task design, we still encourage you to participate in SENSEVAL-2.

The following steps are involved in designing/organizing a SENSEVAL-2 task:

1. Choose the language.
2. Choose the task.
3. Choose the dictionary (sense inventory) and corpus.
4. Prepare the test collection:
    - Choose a lexical sample (depends on the nature of the task).
    - Have the test collection manually tagged (the 'gold standard').
    - Divide test collection into a training set and an evaluation set.
5. Release the data in stages: trial, training, and evaluation.
6. Score the system answers and return the results.

NATURE OF THE TASK

'All-words' vs. 'Lexical-sample' vs. 'In a system'

In the all-words task a participating system will be evaluated on its disambiguation performance on every word in the test collection. In the lexical-sample task a system will be evaluated on a preselected set of words 'unknown' to the system prior to the evaluation. However, the system must nevertheless tag every word that it can, as if participating in an all-words task. The test collection ideally consists of a large corpus of running text. The advantages of the lexical-sample approach are as follows:

1. Quicker development of substantial evaluation materials.
2. Reduces work of preparing training data, because it's required for only a subset of the sense inventory.
3. More systems can participate because a system designed for all-words can participate in lexical-sample, but not vice versa.

However, an all-words task is arguably the more natural and useful application of a WSD system, for which evaluation results are more interpretable.

In the third type of task, a WSD system is evaluated in the context of a real NL system such as MT, IE, etc. In this task, a WSD system is evaluated on how well it improves the performance of the NL system. While this may be the most beneficial way of evaluating WSD systems, it is also the most difficult type of task to design. Every participating WSD system would have to interface with the same NL system. If it isn't possible to evaluate within a real NL system, a lexical-sample task can be designed instead in which the sense inventory is carefully chosen to reflect sense distinctions required of the NL system (for instance, choosing words whose senses that are lexicalize differently in different languages.)

DICTIONARY AND CORPUS

There are two main issues in choosing a dictionary, or sense inventory. First, the definitions or descriptions of the senses should be 'full' enough both to support consistent manual tagging and to provide good input to those systems that rely on definitions. Second, it is best if no previous WSD work has used the sense inventory, so no particular WSD system has an advantage. (Despite these concerns, WordNet and/or EuroWordNet have been put forward as candidates for English SENSEVAL-2, because of their availability and coverage.)

The corpus should be diverse and balanced, reflecting the kinds of texts that are used in real NL applications. It should contain a good mix of all the senses in the dictionary. Of course, availability for research purposes is a critical issue, since the corpus will have to be distributed to all of the participants. Depending on the task there are additional considerations:

Lexical-sample: The corpus should be as large as possible given the above constraints: we suggest a target of 100 instances of each word to be evaluated. The number depends on the complexity of the disambiguation task for each word. More difficult words (roughly, more frequent and more polysemous words) might require more instances. Also, since some WSD systems use a large context surrounding each instance, it is preferable to provide as much context as is possible. Running text is best; but, in the minimum, at least two sentences on each side should be provided.

All-words: The corpus should be at least 5000 words long, preferably longer. It should contain a representative mix of words and word senses. The length is constrained only by time available for manual-tagging (see below for estimates).

CREATING THE TEST COLLECTION

Selecting a lexical sample

A suitable cross-section of words has to be chosen from the whole dictionary. The total number of words chosen is constrained only by the time available for manual-tagging (see below for estimates). There is no accepted method to choose a suitable inventory, but the following are several criteria to consider:

- Part of speech of the word.
- Frequency of the word.
- Polysemy of the word.
- Distribution of the word's senses (uniform or skewed).
- Other measures of WSD difficulty.
- Type of polysemy (homograph, major sense, regular polysemy, has various translations, etc.).
- Selection in previous or current SENSEVAL tasks (even across languages).

Use your criteria to reduce the dictionary to a set of candidate words, and from this take a random sample. For example, in English SENSEVAL-1, sample words were allocated to the cells of a 4×4 grid of frequency vs. polysemy. Words were also chosen to cover three parts of speech (nouns, verbs, and adjectives), with some words having multiple parts of speech.

Manual tagging of the corpus

A manually-tagged corpus provides the 'gold standard' to which all participating systems will be compared, thus its preparation is critical to the whole evaluation exercise. A gold standard corpus must be replicable. To prove this, several human taggers should tag the corpus. If the taggers agree in most cases (at least 90%), then the corpus is considered replicable. The level of inter-tagger agreement can also be used used as an upper bound on expected system performance.

Every word should be tagged by at least two human taggers. It is not necessary to tag every instance of a word: as mentioned above, 100 instances is a reasonable target. Taggers should work succesively through the chosen instances of a word, using concordance software, rather than move word-by-word over the whole corpus. Multiple tags should be allowed. If the human taggers disagree, then some form of arbitration is required. In English SENSEVAL-1, for example, a third human tagger chose which tag(s) should apply. Alternatively, the taggers could reach a consensus on their own. A less satisfactory scheme is to take the disjunction of all tags. The arbitrated corpus is the gold standard.

For each word, compute inter-tagger agreement using individual taggers against the other taggers *before* arbitration. There are many ways to compute inter-tagger agreement. Bruce and Wiebe (1999) analyse the use of *Kappa* statistics. You could also use your answer scoring method. (Scoring methods are not discussed in this document.) An average level of agreement (over all words and all humans) of at least 90% is desirable.

Cyclical development Developing the sense inventory and annotated corpus can be done cyclically. Annotating the corpus might expose problems in the sense inventory in either untenable sense distinctions or poor descriptions of the senses. If time allows, the sense inventory should be revised, and the corpus retagged, as many times as is necessary. Of course, this should be done very carefully, so that the sense inventory does not come to reflect only the small corpus sample you are working with.

Other annotations Corpus instances can also be tagged with part of speech, WSD difficulty scores, or any other type of annotation. In SENSEVAL-1, difficulty scores were not assigned, but it would have been very useful if they *had* have been assigned, because it would have allowed more tenable comparisons between the systems that disambiguated only subsets of the dataset.

Human taggers Arguably, it is preferable to have expert taggers (i.e., professional lexicographers) do the work, because the difficulty of tagging would seem to require linguistic and lexicographic expertise. However, experts are expensive, so if non-experts can do the same quality of work, they could be used instead. Whether non-experts can tag with the same proficiency as experts will be tested in a pilot study associated with the development of the English resources for SENSEVAL-2.

Time estimates Experience from SENSEVAL-1 suggests that it takes one person one minute to tag one instance of a word in a corpus (assuming all instances of the same word are tagged in succession). Two taggers may disagree as much as 40% of time requiring arbitration, which also takes about one minute. So for example, having two taggers tag 100 instances of 40 words would take 2×4000 minutes plus 0.4×4000 minutes, giving 9600 minutes = 160 hours.

Format

The format of the data collection should be fixed early in the process. It should be kept as simple as possible with a minimal use of SGML-style tags. It should not be changed after the trial data release.

TEST COLLECTION RELEASES

We suggest that you prepare three separate releases of data.

Trial data release The *whole* corpus, a portion of which is annotated with 'dummy' annotations, and the format of the dictionary entries. The whole sense inventory itself may be released, with a proviso that the content—not the format—might change. The purpose of this release is twofold: to freeze the data formats, and to provide training data for the (unsupervised) systems. Obviously, the lexical sample should not be released.

Training data release The finalized sense inventory and a portion of the gold standard (annotated) corpus. The purpose of this release is to provide training data for supervised systems. We suggest that half of the gold standard corpus be released, retaining the other half for evaluation. Unsupervised systems can ignore this release (except for the sense inventory). Before receiving the data, participants must agree to freeze their code.

Evaluation data release The other portion of the corpus, with annotations (i.e., the answers) removed. Systems should be instructed to tag every word in the corpus with one or more senses of the inventory, and to return answers in a specified format.

This scheme is different from Resnik and Yarowsky's (1999) proposal, which did not permit the test words to be released before the evaluation. Their scheme does not allow supervised systems to participate (at least in the first year that such an evaluation is run).

SCORING

Scoring involves designing baseline algorithms and scoring methods. The answers from every system must be scored and the results returned to the participants. Then the systems must be compared, and conclusions drawn in preparation for the SENSEVAL-2 Workshop. This document does not discuss the details.

REFERENCES

Bruce, Rebecca F. and Janyce M. Wiebe. Recognizing subjectivity: a case study in manual tagging. *Natural Language Engineering* **5**(2):187–206.

Calzolari, Nicoletta and Ornella Corazzari. Senseval/Romanseval: The framework for Italian. *Computers and the Humanities* 34(1-2) (Special Issue on SENSEVAL).

Kilgarriff, Adam and Joseph Rosenzweig. 2000. Framework and Results for English SENSEVAL. *Computers and the Humanities* 34(1-2) (Special Issue on SENSEVAL).

Resnik, Philip and David Yarowsky. 1999. Distinguishing systems and distinguishing senses: new evaluation methods for Word Sense Disambiguation. *Natural Language Engineering* **5**(2):113–134. (`http://www.cs.jhu.edu/~yarowsky/pubs/nle00.ps`)

Segond, Frédérique. 2000. Framework and results for French. *Computers and the Humanities* 34(1-2) (Special Issue on SENSEVAL).

SENSEVAL website: `http://www.itri.brighton.ac.uk/events/senseval/`