Writing Annotation Instructions

Janyce M. Wiebe
Dept. of Computer Science and the Computing Research Laboratory
New Mexico State University
wiebe@cs.nmsu.edu

September 17, 1997

1 Introduction

This document reports the results of the *Writing Annotation Instructions* working session of the "Tagging Text with Lexical Semantics: Why, What, and How?" workshop sponsored by the Association for Computational Linguistics, Special Interest Group on the Lexicon. The workshop was held in conjunction with the 1997 Applied Natural Language Processing workshop, 4-5 April 1997 in Washington D.C.

I lead the working session, and the participants were the following:

- Allen Browne
- Martha Evans
- Clare Voss
- Hwee Tou Ng
- John Lowe
- Adam Kilgarriff
- Christiane Fellbaum
- Patricia Robinson
- Alain Polguere
- Janyce Wiebe

The participants have a wide range of relevant experience. The working session resulted in a number of useful suggestions.

This is a draft document. For future reference, attributions and notes about future refinements are included in parentheses. The reader can ignore these parenthetical comments.

2 Defining the Annotation Set

Although not the focus of the working session, defining the annotation set naturally came up in discussion. (Note to all: I do not have very much in my notes about this section. Suggestions concerning what to add are welcome. Perhaps this should not be a separate section, but rather merged in with the second proposal below in section 4.2.)

There was some discussion of the generative lexicon approach, contrasted with listing individual meanings (Adam and Christiane; is there some discussion you would like to include here?)

It was suggested that meaning hierarchies are not the same between different languages; one cannot take the lexical tree developed for one label, change the labels, and use it for another language (Christiane and Alain; do we want to pursue some discussion of multi-lingual tagging?)

It was pointed out that dictionaries are not written with the same purposes for which we use them in computational linguistics. The dictionary is designed to aid human language users. Including many senses may help a writer, by presenting greater variety. Redundancy among the senses does not make the dictionary less usable to writers. On the other hand, if dictionary entries are treated as sets of mutually exclusive, exhaustive categories, as they often are in training and evaluating automatic word-sense disambiguation systems, then fewer senses are better, and redundancy is undesirable. (I remember this point being made by Christiane and Allen; Christiane remembers it being made by Adam.)

It was also pointed out that lexicography is evolutionary, whereas, to the extent possible, we want the tagsets for training and evaluating automatic systems to be fixed (Adam, Alain).

The issue came as to whether the annotations should be at the word-sense level or at a higher level of groupings of related senses. It was noted that such groupings are available in WordNet 1.5 for nouns (Hwee Tou), and will be available for other parts of speech in WordNet 1.6 (Christiane). Such groupings will provide a smaller set of non-redundant classes.

The question came up how much feedback should be elicited from the taggers concerning the tagset. For example, should the instructions direct them to document what they perceive as errors in the instructions, as well as missing, redundant, and incorrect senses? In a working session lead by D. Yarowsky and P. Resnik at the same workshop, a proposal was developed to produce a shared corpus for evaluating word-sense disambiguation algorithms. During the session, a step was added to the proposal to revise the tagging instructions annually, to take advantage of feedback from the taggers.

In her study presented at the workshop, Christiane and her colleagues found that ordering the senses randomly makes the tagging task more difficult for the annotators, but that doing so leads to more "scrupulous examination of the available choices." (Fellbaum et al., p. 39) Although it would involve more effort, it would be wise to order the senses randomly to obtain better annotations. As discussed below, participants felt that the more the annotators know about the categories and task, the better, so it was suggested that one should be up front with the taggers about the ordering of the entries.

Many participants in the group have experience tagging distinctions other than word-sense data, or have plans to do so in the future (Clare, Patty, Martha, John, Jan). In these cases, defining the tagset is often even more open-ended than for the word-sense problem. (If there is interest, we can elaborate more on the differences in developing instructions for word-sense tagging and for other kinds of tasks. For example, there are often representational issues with other tasks. For word-sense tagging, the taggers are simply putting objects in boxes; for other tasks, such as reference resolution, the annotators must indicate the links between objects in a uniform way.)

3 Choosing the Annotators

Many felt that requiring the annotators to be native speakers is too strict. Annotators who attend college in the target language, for example, especially those studying an area such as linguistics, may be appropriate. If they attended school in English, it is likely they learned their world view in English (Martha).

Should the taggers be experts or naive? (Christiane) To some extent, this depends on the purpose of the annotations: for developing an ontology to support word-sense disambiguation, or for identifying subcategorization frames to include in a lexicon, one is likely to use people with expertise in linguistics. But to evaluate an NLP system, one may want to use more naive annotators. (Jan; any other discussion?) An interesting question is, do we want the naive annotators to agree with one another? Or do we want their tags to agree with the experts? (Christiane)

A middle ground is to choose annotators without background in linguistics, lexicography, or computational linguistics, but make them pass a test first. The test could determine that they have good language skills, are precise, and are able to pay close attention to detail. This seems appropriate for the kind of significant effort proposed here. (Jan)

4 The Overall Process

The obvious question came up: is it possible for annotators to perform word-sense tagging with high inter-coder reliability? Some participants expressed doubt, based on what has been achieved so far (Christiane; Martha). However, an extended process such as the proposed below has not yet been attempted on a large scale.

One general point of agreement is that, to have any hope of achieving high inter-coder reliability, an annotation project requires a significant amount of time and effort. It is not realistic to expect to give a short set of instructions to taggers without training and have them reliably annotate the corpus. It might take a month or more of effort for a tagger to become proficient. (My notes attribute the month figure to Alain, but I did not record exactly what he meant to include in that time: the instructions as well as training of the annotators? Just training, once the instructions have already been written? Do we want to attempt to give general projected time requirements?)

It was suggested that taggers should be asked to tag a word one word at a time, i.e., to perform "targeted tagging" (Christiane; this was also suggested in the Yarowsky and Resnik working session). Targeted tagging was employed in tagging the DSO corpus developed by Hweetou Ng and recently made available through the Lexical Data Consortium.

At the start, it should be determined what is a reasonable level of agreement (Patty).

The following proposals emerged during the course of the session. (Comments are welcome!) The first is a more high-level view of the process and the second is more detail. The proposals are compatible.

4.1 Proposal 1

(This is Clare's picture).

```
 I.  Develop Lexicographic standards
     ''Analysis''

            |
            |
            V

II.  Develop Tagging Instructions and materials
            |
            |
            V
III. Training of Annotators
            |
            |
            V
IV.  Testing/Evaluation of Annotators

Revise the standards (go back to I) or the tagging instructions
  and materials (go back to II).
```

This proposal is for a multi-stage process. In the earlier rounds of tagging, let the annotators mark any ambiguities they find difficult. Then, one can revise the instructions to address the worst ambiguities found (such as grouping or paraphrasing the categories/senses; see step 3 in the next section).

4.2 Proposal 2

Many participants suggested that the more the taggers understand about the task, the better their annotations will be.

This includes helping them understand how the tagset was developed and how the annotation task was defined (Adam, Alain).

Step 1:
The taggers should work with the problem first, to appreciate how difficult it is (Alain). That is, they should be given a tagset, some data, and only rudimentary instructions, and then have a go at trying to perform the task.

This will show them how much variation there is in the data and how many issues arise in the problem.

Step 2:
Give the taggers a conceptual picture of the distinctions being made (Alain, Adam). Explain the end-application and purpose of the annotations. This will help them realize why they must be so consistent. In addition, they may be more productive if they feel they are contributing to a concrete project (Patty).

Adam made a number of concrete suggestions concerning what kinds of meanings there are and how they relate to one another, sparking lots of good discussion.

```
                    Homonymy


Analogy             (Polysemy)           Alternation
                                      (Regular polysemy)

(Hierarchy)

                    Collocation
```

The idea of the picture is that a particular instance can be classified as a mixture of the various categories. That is, an instance can have elements of each. Polysemy is the interaction of the four things in the picture.

Following are some suggestions and discussion of useful conceptual classes. (Corrections are welcome; perhaps below is not the intended categorization.)

Regular polysemy: patterns that reoccur often, as in "I'll have a whiskey". This includes metonymy.

Collocation: a pattern you have learned already, such as "for some time", "see the light", "see the light of reason". The last two show that such patterns are not always fixed, but may involve a grammar (this was also suggested by an audience member during the workshop). (I wrote down: "any kind of multi-word unit". Do we want that comment here?)

(Alain:) we should distinguish between real lexical units and those that are not. For example, "ask questions" is a lexical unit, but, e.g., "kick the bucket" is not (Martha).

There was disagreement among the participants on this issue. For the purposes of the working session, various conceptions are possible, so a policy should be formed to use one treatment or the other, to tell the tagger what you want them to do. (A number of participants made this last point).

What level unit should the taggers tag? Are they to tag “take”? Or, “take part” (and not tag the individual words “take” and “part”). Or, “ask a question”? This decision will influence what conceptual picture is used, and, obviously, the policies you form. A suggested policy is to include whatever multi-word units you wish to be tagged in the dictionary, and then ask the annotators to tag the largest unit for which there is an entry (Jan).

Another class for which a policy needs to be developed is analogy, as in “She had a few spots missing off her dice.” These cases are distinct from regular polysemy and collocations, but represent a novel combination of the two.

Where does conventional metaphor belong in the conceptual picture?

Christiane suggested that her conceptual picture is different. The phrases “top of the mountain” vs. “top of the head” vs. “top of the garment” all have different meanings. They are not homonyms and are not instances of regular polysemy. Where do they fit into the picture?

It was at this point that Adam added “Hierarchy” to the diagram, thus adding another dimension to meaning: generality versus specificity.

Step 3:

Step 2 above concerns making the taggers aware of problem areas and distinctions that the taggers will run across in the data. The next step is to develop policies for dealing with these issues, and expressing them clearly for the taggers. The process is complex: the taggers must first understand the sentence itself in terms of the conceptual picture developed in step 2. Then they have to understand how to map their understanding to the dictionary (Christiane).

Because the tagger has experienced the difficulty of the problem, and understands some of the complexity involved (from Steps 1 and 2), he or she will be able to understand that these are policies adopted to make tagging feasible, and are not the only possible rules that could have been used. Explicitly acknowledging cases in which alternative policies are possible will assist the tagger. They will perceive more than one interpretation in various occasions, and we want them to realize that this is a problem inherent in the task, not in how they are performing it. (Jan)

(Alain:) In addition to developing policies for conceptual complexities in the problem, one has to develop policies for how to use the dictionary. One has to analyze how well the dictionary distinguishes among the word senses, for the purpose at hand, and develop instructions for using it. If some dictionary is being used that does not follow your lexicographic theory or methodology, then you have to create an interface between your decisions the dictionary.

(Alain:) For example, one might find it necessary to paraphrase some senses. Or, one might find a set of senses that are too close to distinguish for the purpose at hand. In this case, it could be specified that there is just one sense corresponding to the set of senses.

This step also involves specifying which of the dictionary's resources are to be exploited. (Christiane:) For example, there may be groupings of senses already present in the dictionary. Should the annotators tag at the higher level of sets of related senses? It was noted that such groupings are available in WordNet 1.5 for nouns (Hwee Tou), and will be available for other parts of speech in WordNet 1.6 (Christiane). Such groupings will provide a smaller set of non-redundant classes. In addition, one is not intended to use a dictionary like WordNet by considering the senses in isolation. The taggers should be directed to consider the synonymns of the senses and the position of the sense in the net (Christiane).

Note that the above concerns giving the taggers pre-formed groupings of particular sets of senses to use. Another issue to address at this stage is the degree of grouping of senses that the taggers should be allowed to do. One suggestion is that the taggers be allowed to tag tag up to two possibilities (Christiane). Some restriction is probably wise. Arbitrary groups might made using annotations to evaluate automatic classification systems too difficult (Jan).

These policies form part of the instructions.

Step 4:
The annotators should read the instructions and training materials, and request and receive feedback. A suggestion was made that this be done electronically, so there is a written record (Martha, Jan).

Next, the annotators should annotate the training data with respect to the instructions. One part of this phase should involve gathering the the annotators together in one room and tagging the training data together. This process will bring out questions that individuals may not have thought of on their own (Allen, Martha). In addition there could be a part of this phase in which experienced annotators work with the inexperienced ones (Allen). Another possibility is for the annotators to annotate the data on their own, then perform a "diff" on the files, and talk about the differences in their decisions (Patty). To the extent feasible, a written record of the questions and decisions during this phase would be very useful (Martha, Jan). In any event, the annotators should receive feedback on their annotations before going on to the next phase.

The annotation–feedback process can be iterated until the annotators feel comfortable and their performance on the training data is sufficiently accurate (measuring their answers against more expert annotations).

To perform an inter-coder reliability study, the annotators are given fresh test data and are not allowed to receive any further feedback.

5 Training Materials

The training materials should include sample annotated documents, and the instructions should contain examples of all of the issues discussed (Clare). Experience has shown that taggers consult the annotated documents (training data) very frequently (Patty).

The question arose as to how much training data is appropriate, per word, for training human annotators to perform word-sense disambiguation? (Adam; do we want to try to suggest a size?)

It was suggested that, to the extent feasible, the questions and answers in the training process should be written down (Martha, Jan). These are very important resources for training and for recording the decisions made over time. The written results from training one set of taggers can be incorporated into the training materials for the next set of taggers, so the taggers can see what questions the previous taggers had and the answers they give. One idea is to mix up these questions and answers in the training material so that difficult and easy questions, and questions concerning various issues are mixed in together. The idea is that mixing up the questions will provide better exercise to the taggers.

In addition to samples of already tagged data, a list of contexts in which a sense has been used can be a useful part of the training materials. For example, some example uses can be displayed with the dictionary definition, as in the following:

```
Sense 4
market, securities industry -- (the securities markets in the
aggregate; "the market always frustrates the small investor")
       => industry -- (the people engaged in a particular kind of
          commercial enterprise)

market sentiment, the market again showed little interest,
```

`major market activity, fundamental market conditions, the 1980s bull market, market researcher, yesterday's market report, the market drove the dollar downward, current market conditions`

Based on his experience in preparing the DSO corpus, Hwee Tou suggests that the existing sense definition glosses in WordNet give insufficient details for human taggers to adequately distinguish among the refined sense distinction to achieve a high enough level of inter-tagger agreement. Thus, in addition to the kind of listing given just above, he suggests that actual examples showing the usage of a word sense in context are very helpful for human taggers to help decide on the ambiguous cases (Hwee Tou)

A number of participants recommended giving the taggers general strategies to follow, such as the following:

- Working by analogy, i.e., considering the similarity of the instance to previously tagged instances. (Patty; shall we elaborate on this point?) (Alain, you also had some comments in this area concerning lexicographers versus taggers, but I'm afraid my notes are too cryptic on this point.)
- This item is a subcase of working by analogy: considering the collocations that senses appeared in in previously tagged instances. However, some participants suggested that the taggers should be warned not to assume that collocations uniquely determine word sense.

6 References

Fellbaum, Christiane, Grabowski, Joachim, and Landes, Shari, Analysis of a hand-tagging task. In *Proc. ANLP-97 Workshop, Tagging Text with Lexical Semantics: Why, What, and How?*, Association for Computational Linguistics Special Interest Group on the Lexicon, pp. 34-40.