Last Words

Mark-up Barking Up the Wrong Tree

Annie Zaenen*
PARC

The interest in machine-learning methods to solve natural-language-understanding problems has led to the use of textual annotation as an important auxiliary technique. Grammar induction based on annotation has been very successful for the Penn Treebank, where a corpus of English text was annotated with syntactic information. This shining example has inspired a plethora of annotation efforts: corpora are annotated for 'coreference', for animacy, for expressions of opinions, for temporal dependencies, for the estimated duration of the activities that the expressions refer to, and so on.

It is not clear though that these efforts are bound to repeat the success of the Penn Treebank. The circumstances in which the Penn Treebank project was executed are vastly different from those in which most annotation tasks take place. First, the annotation was a linguistic task and one about which there is reasonable agreement. People might quibble about the way to represent certain constituent structure distinctions in English, but they do, in general, not disagree about the distinctions themselves; and if you don't like the Treebank as is, you can translate it into your favorite format. Second, the work was done by advanced students who understood the task and were supervised by specialists in the field. Third, this was not done in a hurry. The project started in 1989 and the corpora are still maintained and the annotations improved. About the only thing that this project has in common with the bulk of annotation tasks is that the annotators were not very well paid.

Currently, we see annotation schemas being developed for phenomena that are much less well understood than constituent structure. In workshops and conferences we hear lively discussions about interannotator agreement, about tools to make the annotation task easier, about how to cope with multiple annotations of the same text, about the development of international standards for annotation schemes in specific subdomains, and, most importantly, about the statistical models that can be built once the annotations are in place. One thing that is much less discussed is whether the annotation indeed helps isolate the property that motivated it in the first place. This is not the same as interannotator agreement. For interannotator agreement, it suffices that all annotators do the same thing. But even with full annotator agreement it is not sure that the task captures what was initially intended. Assume that I want to mark all the entities in a text that refer to the same entity with the same number and I tell my annotators "Whenever you see the word Chicago, give it the same number": I'll get great interannotator agreement with that guideline but it is debatable whether I will realize my proclaimed aim of classifying references to one and the same entity in the outside world. Presumably, I would like to catch all the references to the city of Chicago, but Chicago pizza is made and sold all over the United States, and the relation

* Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304, USA; zaenen@parc.com.

between The University of Chicago and the city of Chicago is about the same as that of the DePaul University and the city of Chicago.

The problems with the ‘coreference’ annotation tasks of MUC and the like are well documented and not solved. Kibble and van Deemter (2000), for instance, discuss the difficulties created by the assumption that coreference is an equivalence relation, and hence transitive for cases like the following: *Henry Higgins, who was formerly sales director of Sudsy Soaps, became president of Dreamy Detergents. Sudsy Soaps named Eliza Doolittle as sales director effective last week.* Given the way coreference is defined, Henry Higgins is coreferent with Eliza Doolittle. The persisting discussion about how to define the task (Is it better to try to get to coreference or anaphoric dependency? Is the former part of the latter?) shows that we are not confronted with a trivial problem. There is no reason to assume that the other annotation tasks that are currently undertaken are all simpler in nature.

There are two aspects of annotation tasks that make them more difficult than they might seem at first blush. The first is inherent in the kind of annotations that are currently needed. The field is moving from information retrieval to language understanding tasks. To understand a linguistic utterance is to map from it to a state of the world, a non-linguistic reality. Language understanding always has a non-linguistic component. In computational settings, unfortunately, most of the time we do not have independent access to this non-linguistic component. This means that language understanding systems have to be more than just language understanding systems: One expects them also to take care of some minimal understanding of the world the language is supposed to describe. But relations between linguistic entities and the world have not been studied in linguistics nor anywhere else in the systematic way that is required to develop reliable annotation schemas: Traditional formal semantics says how meanings are put together and remains silent about semantic primitives. Lexical semantics is very fragmented: One part of it tends to limit its scope to lexical items that exhibit syntactic alternations, whereas another part concentrates on improving traditional lexicography.

There certainly are efforts that are relevant to the study of the relation between language and the world, but they are the proverbial drop in the ocean. The kind of work needed falls between artificial intelligence and computational linguistics and is treated as peripheral to both. To illustrate this again with the coreference example: In linguistics there is exquisite work on the interpretation of reflexives and reciprocals, but they account for less than 5% of pronouns. There are influential theories on the interpretation of definite and indefinite noun descriptions, but a naïve interpretation of the currently most influential of these theories, the novelty/familiarity theory, leads people to the wrong expectations about the status of definites as documented by Poesio and Vieira (1998). There is work on quotations; there is work in psycholinguistics; but all this doesn’t add up to a coherent account of the use of referential expressions in text and dialogue.

Annotation tasks typically involve these ill-understood phenomena. Current practice seems to assume that theoretical understanding can be circumvented and that the pristine intuitions of nearly naïve native speakers can replace analysis of what is going on in these complicated cases. The results that I have looked at suggest that this is wishful thinking. To take an example that looks deceptively simple: When does a linguistic expression refer to an animate being? In *I went to the restaurant on 10th street. {It makes|They make} great clam chowder*, the restaurant seems to be a place, but the *they* seem to be people and with *it* we seem to be referring to an organization. One can put together a coherent picture about how we understand such sentences but it is not trivial and it does not translate immediately into a quick and unambiguous annotation scheme. Annotators either vacillate or they interpret the guidelines in a reductive manner; for example, anything that follows directional *to* is a place and anything that is anaphorically related to a place is also a place. The problems are not limited to animacy annotations: Questions such as 'What does it mean to detect the point of view of the author of a text?' and 'Is he supposed to have one point of view toward everything he talks about?' are not easier to answer, as illustrated by a recent discussion in the Aquaint project. A pilot evaluation of opinion annotations produced very low F scores and the authors were at loss about how to show the advantages of such annotations. Corpus-based research risks becoming as bedeviled by annotation problems as theoretical linguistics is by the questionable methodological status of grammaticality judgments.

The second problem that annotation tasks face is not inherent. It is created by the current funding model. In the name of accountability, current NLP practice is wedded to quantifiable results, short time horizons, and strict financial control. In this setup there is no time for fundamental research. So, when research is necessary, it has to be called by another name. Part of it will be shuffled under the heading 'annotation'. But narrowly defined annotation has nothing to show for itself: It is an auxiliary preparatory task destined to facilitate others. The annotation work has to be done well, one assumes, but, surely, it must be done quickly. The often subtle distinctions involved need to be made by annotators with little training, and only a minimum of time can be spent on defining the task and testing the guidelines. If annotation is made into a separate project, the schedule remains tight, and one has to argue that the same annotation will be useful for various applications to get sufficient funding. This carries the additional danger that annotations intended to serve for several, often not yet defined, applications may in fact not be useful for any. However the task is organized, annotators typically have no stake in the end result. If one has to slip research under this heading, the task becomes impossible.

Of course, as long as the task is to provide material to develop and refine machine-learning techniques, much of this doesn't matter. Whether *Henry Higgins and Eliza Doolittle* are referring to the same entity or not is of no interest in that context. The technique has only to show that if it is told that they are coreferent because they had the same job (even at different moments), then it can also learn that George Bush and his father are coreferent. Whether that is the way we understand natural language is not what the techniques are about. The problem arises only when one starts to use the results in further processing or tasks such as question answering.

Given the inherent difficulties of annotation tasks and the lack of appropriate funding to do the research that would allow them to be done well, one might wonder whether one could avoid the issue completely. There certainly are researchers who think that every NLP task can be recast as a text-to-text mapping task and all relevant regularities can be found through appropriate alignments. It is not obvious that this will work for purely linguistic regularities, and when the task is to discover properties of a non-linguistic reality, it is clear that not everybody is willing to bet on it.

Thus for the foreseeable future we are stuck with annotation. We should recognize that annotations are no substitute for the understanding of a phenomenon. They are an encoding of that understanding. The encoding is different from a rule-based encoding in that it does not require a generative formalization and it allows a more piecemeal approach. It is reasonable to ask that the understanding of (some aspects of) the language–world mapping be encoded in annotations. It is also reasonable to ask that an important part of the fundamental research done in this area be embodied in annotations to ensure that the research remains linked to phenomena that occur in real text and dialogue, but a serious investment needs to be made to understand the mapping between language and the world itself better. To call the study of these problems "development of annotation guidelines" belittles their scope and importance.

References

Kibble, Rodger and Kees van Deemter. 2000. Coreference annotation: Whither? In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pages 1281–1286, Athens, Greece.

Poesio, Massimo and Renata Vieira. 1998. A corpus-based investigation of definite description use? *Computational Linguistics*, 24(2):183–216.