Using Volunteers to Annotate Biomedical Corpora for Anaphora Resolution

S. Watters, B. McInnes, D. McKoskey, T. Miller, D. Boley, M. Gini, W. Schuler
Dept of Computer Science and Engineering, University of Minnesota

A. Polukeyeva, J. Gundel
Dept of Linguistics, University of Minnesota

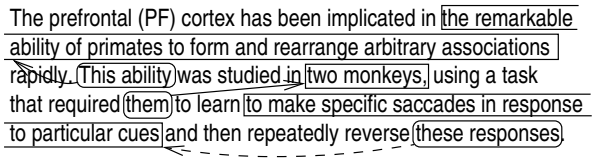
S. Pakhomov, G. Savova
Division of Biomedical Informatics, Mayo College of Medicine

Abstract

The long-term goal of this project is to build an annotated corpus of biomedical text, to be used as a foundation for the development of automated anaphora resolution systems. We plan to explore the feasibility of using a community of volunteers to annotate a corpus drawn from publically available biomedical literature. We present issues in creating such a community and discuss results obtained from a pilot study.

A motivational example

In the past several years, research in machine processing of natural language has seen a resurgence of interest in issues involving anaphora resolution, the process of determining the intended referent of phrases whose interpretation depends on prior linguistic context. For example, consider the following text:



This text contains three anaphoric expressions: "this ability", "them", and "these responses". One of these ("them") is a pronominal anaphor and the other two are full noun phrase anaphors whose anaphoric function is signaled by the determiner "this/these".

The first two anaphoric expressions, "this ability" and "them" must be interpreted as having the same reference respectively as "the remarkable ability of primates to form and rearrange arbitrary associations rapidly" and "two monkeys". This type of anaphoric expression, where there is a preceding noun phrase with the same referent (coreference), is the most common. The third example, "these responses" is not coreferential with a preceding noun phrase. It must be interpreted as referring to the monkeys' responses to the cues in the task, which can be inferred from the full clause "This ability was studied in two monkeys, using a task that required them to learn to make specific saccades in response to particular cues."

The problem

Interest in anaphora resolution has been driven by the need for natural language processing (NLP) systems that can handle a variety of tasks, including information retrieval, information summarization, information extraction, language generation and understanding, and machine translation of natural language. Although the area of research addressing anaphora resolution issues is far from being in its infancy, there are still a number of outstanding methodological issues that need to be addressed. One such issue is how to create corpora for developing and evaluating anaphora resolution algorithms, as there are currently no set guidelines for annotation or formatting. The need for such guidelines has been recognized by other researchers. For example, as Salmon-Alt & Romary (2004) note, "There is an opportunity to stabilize the corresponding knowledge as an international standard in the context of the recently created ISO committee TC37/SC4 on language resource management. Indeed, this committee aims at providing generic standards for the representation of linguistic data at various levels."

Our goal is to develop a methodology and a framework for manual annotation of textual corpora for anaphoric relations, and to apply this methodology to build an annotated corpus of biomedical text. Such a corpus would serve as a foundation for the development of automated anaphora resolution systems for the purpose of extracting searchable semantic content. This requires development of a uniform annotation system for collecting data for training, evaluating, and reliably comparing results across systems. As part of this project, we address the feasibility of using a community of volunteers to annotate a corpus drawn from the publically available biomedical literature in the form of published papers and abstracts.

Related work

Most researchers prior to 1985 handcrafted data sets to evaluate their algorithms. For example, Hobbs (1978) used three texts (William Watson's "Early Civilization in China" pp 21-69, Arthur Haley's novel "Wheel", pp 1-6, and the July 7, 1975 edition of "Newsweek", pp 13-19) and Lappin & Leass (1994) used a corpus of 82,000 words obtained from five computer manuals. After 1985, it became more common for researchers to use readily available data

such as MUC-6 (Message Understanding Conference) and MUC-7 data (Soon, Ng, & Lim 2001; Ng & Cardie 2002; Yang *et al.* 2003), TRAINS93 dialogues (Strube & Müller 2003), and Medline literature (Castaño, Zhang, & Pustejovsky 2002).

Over the years, researchers have realized that publicly available corpora are not only desirable but are necessary to ensure that the evaluation of an algorithm's results can be interpreted in a reliable manner and that the measures used to determine an algorithm's success are consistent. Many issues involved in annotation are similar regardless of the annotation task, for example how much data is enough, what exactly needs to be annotated, and how to evaluate agreement. Due to these types of issues and in order to create large corpora in a reasonable amount of time, a number of research projects have begun addressing how volunteers can be used for annotating corpora. For instance, Mihalcea & Chklovski (2004) describe a project where volunteers provide tagging using a web-based application "that allows contributors to annotate words with their meanings." Although the tagging is for word sense disambiguation and not anaphora, the project addresses many of the same issues that our own project faces. Other applicable areas of research are validation techniques such as described in Chklovski & Mihalcea (2003) for validation of the data collected from the volunteers. This is an important problem that needs to be addressed to ensure that only correctly tagged documents are included in the corpus.

Some challenges and issues, however, are unique to the task of (co)reference annotation. These include, for example, whether or not annotators need to be domain experts (in our case, experts in biology/biomedicine), how to elicit reliable information about the precise referent of an anaphoric expression, and what is the best format for annotation.

There is a rich literature on algorithms for coreference resolution (see Mitkov 2002 and the Proceedings of DAARC, Discourse Anaphora and Anaphora Resolution Conferences), including some work in the biomedical domain (Castaño, Zhang, & Pustejovsky 2002). Preliminary work is under way to develop an on-line system for evaluation of coreference resolution (Popescu-Belis *et al.* 2004), but we are not aware of any effort on the scale of the project we describe here, in terms of the size of the corpus we are aiming at annotating, the range of anaphoric expressions we expect to include, and the type of documents we are annotating. Since our focus is on published articles in the biomedical domain, the type of data that must be used for development and evaluation is often difficult to understand for someone who is not a domain expert, and it may be necessary to use volunteer domain experts for the annotation.

Our application shares many of the issues common in many online communities such as assuring that the roles are well defined for both the transient and permanent participants, in such a way that the value of the contributions from each participant is clear (Butler *et al.* 2005; Preece 2000).

Challenges

One of the challenges we face is the integration of an anaphora resolution component into a larger system, which will have a number of preprocessing tasks to perform before the anaphora resolution component will have a chance to process the data. For example, since the anaphoric expressions we are concerned with are noun phrases, the system will require input from a preprocessing module that identifies noun phrases and grammatical features such as number and gender, since these can constrain anaphora resolution possibilities.

There are also other issues that must be addressed before creating a sufficiently large corpus. These include:

1. How much annotation can be done without domain expert annotators?
2. How do we present the data to the annotators and elicit their responses?
3. How do we define "domain expert"?
4. How do we find domain expert volunteers who are willing to assist in annotations?
5. What anaphoric forms do we annotate for (e.g. personal pronouns (them), demonstrative pronouns (this, that), demonstrative determiners (this ability), definite article phrases (the protein)?
6. Do we only include expressions that are coreferential with a previous noun phrase or do we include anaphoric expressions whose interpretation is more indirectly related to previous context (for example 'these responses' in the text fragment above)?
7. How do we determine reliability of the annotation?

The work reported in this paper addresses primarily the first two questions.

Pilot Study

A pilot study was conducted to address two of the issues listed in the previous section, specifically: (1) How much annotation, if any, must be done by domain experts and how much can be done by people who are not domain experts? (2) What is the best way to elicit reference judgments from annotators? The first question is one addressed by any natural language processing project on domain specific corpora. If domain experts are needed to perform the annotation task, then the pool of possible annotators would be much smaller. It may also be difficult to recruit domain expert volunteers since they may not have any "extra" time to give to a project that does not provide them with a tangible benefit.

The pilot study involved the following stages:

1. Selecting an article which represents the types of articles that a volunteer would read and annotate.

   An excerpt consisting of the first four pages from an article from the Journal of Biology 2003, 2:27, entitled "A functional genomic analysis of cell morphology using RNA interference" was selected for use in the pilot study. This journal was selected as representative of the domain that we are investigating.

2. Selecting the type of anaphoric forms to be annotated.

   Since constraints on interpretation of anaphoric expressions differ depending on the form (Gundel, Hedberg, & Zacharski 1993), we restricted the pilot study to pronouns, specifically demonstrative pronouns (*this/these* and *that/those*) and personal pronouns (*she/her, he/him, it, they, its, his, their*). There was a total of 8 pronouns in the analyzed sample, including the forms *those, these, that, their* and *its*,

3. Determining the format for presenting data to annotators and eliciting their responses.

   This addresses the second question posed in the pilot study. As annotators will be asked to identify the interpretation/referent of an anaphoric expression (and possibly also the previous form that it is coreferential with), it is necessary to determine the best format for eliciting this information. The formats considered were "multiple choice" (designed by the investigators) and "free answer." Two main criteria were used in assessing annotation formats for the pilot study:

   (a) Simplicity of Implementation/Scalability
   Questionnaires in free answer format perform much better on this criterion as they can be prepared by simply finding the forms to be annotated and providing annotators with instructions. The multiple choice format, on the other hand, is much more difficult to implement since the choices must be separately prepared for each question, either by a human or by a fairly sophisticated computer program.

   (b) Uniformity/Reliability
   The multiple choice format is more preferable on this criterion as it constrains possible annotations. The free format could be too unconstrained as annotators would have different ways of describing the same referent and/or may use descriptions that do not unambiguously describe the referent, thus making it more difficult to reliably assess inter-annotator agreement. In order to determine whether choice of format would influence the responses provided by the annotators, half were given a multiple choice format and the other half were given a free answer format. The pronouns selected for annotation were identified by having a box drawn around them. The first part of the instruction was identical in both types of questionnaires: "Please read the following excerpt from a published article as if you were reading it as part of your research." The next sentence for the free answer questionnaires was: "For each boxed word, write down what you think the word refers to. If it is not clear, you may write down more than one possibility." For the multiple choice, the instruction was "For each boxed work, select the choice that best represents what the word refers to."

4. Selecting annotators.

   In order to address the question of whether annotators would have to be domain experts, we selected two groups of five for participation in the study; one group consisted of individuals who were domain experts and the other consisted of individuals who were not domain experts. All but one domain expert had a graduate degree in one of the biological sciences (one had only an undergraduate degree). None in the group that were not domain experts had special training in biological sciences. Annotations in free answer format were also carried out by 5 members of our research group. There was agreement on 6 of the 8 questions. Agreement was reached on the remaining two after discussion. The result was included as one of the 5 free answer questionnaires. The reason for having the investigators participate in the pilot study was to address two questions: (i) would linguistic training influence responses/judgments (see Schütze 1996) and (ii) how would responses of investigators compare with those of domain experts?

5. Interpreting the results and drawing conclusions based on these results.

   The results of the questionnaires were compiled and examined. In nearly all of the questions, the domain experts and non-domain experts gave similar or identical answers. To be more precise, the reference problems of the pilot study can be grouped into three categories:

   (a) Those that can be resolved by non-expert majority answer.
   (b) Those that can only be resolved by expert majority answer.
   (c) A small percentage of ambiguous constructions that are resolved incorrectly by both non-experts and experts.

Results and Discussion

The pilot study revealed very little difference between responses from volunteer annotators who were domain experts and those who were not domain experts. The response provided by the majority was the same for both groups in all 8 examples, and there was only one example, question 8, where responses were more consistent for the domain expert group.

Question 8:

> Choosing genes from one chromosomal region would be likely to yield fewer visible phenotypes, whereas choosing genes on the basis of their expression in existing cell lines would assume a correlation between expression levels and function.

All the domain experts responded that the referent of 'their' was genes. Three of the non-experts also thought the referent was genes, but two of them thought it was 'visible phenotypes'.

The observed lack of difference between the two groups is not surprising, given that pronominal reference is constrained by general linguistic knowledge that is independent of subject matter. Thus, it may be that domain expertise becomes relevant only in examples like question 8, where there is true ambiguity. The situation may be different, however, in the case of full NP anaphors, especially 'sortal anaphors' as in 'acetylocholinesterase and butryril cholinesterase....both enzymes', where domain knowledge

may be required to link the phrase 'both enzymes' to the correct antecedent.

If we find similar results in a subsequent study, this could simplify collection of data through distributed systems since domain experts would not be needed in most cases. Annotation of large corpora could be accomplished, for example,as part of a university class in linguistics or computational linguistics that includes reference resolution as part of the subject matter. Students would annotate the corpora as part of the class and the results could then be used in building the annotated corpus. If there was sufficient agreement between non-expert annotators, the results could be used as is, and only indeterminate cases would have to be distributed to domain experts.

The pilot study also provided no evidence of advantage of multiple choice answer format over a free choice fill in format, as the answers provided were the same for both formats in most cases. Also, responses from the investigator group did not differ from those provided by the other free format non-expert or from the domain experts. There were two examples, one from a domain expert and one from a non-expert, where an individual provided a short answer in the fill in format whose extension was clearly more general than the referent of the pronominal form. The domain expert example was a response to question 5.

Question 5:

For example, treatment with a drug that prevents the polymerization of filamentous (F-) actin caused Kc167 cells to develop long microtubule-rich processes, a morphological change similar to [that] observed upon treatment with dsRNA corresponding to the gene encoding Cdc42 GTPase.

One annotator identified the referent of 'that' as 'change', when the more precise response provided by all the other annotators, would have been 'morphological change'. It could be argued that this annotator would not have chosen the shorter answer in the multiple choice format, where 'morphological change' was explicitly provided as a choice. But it is noteworthy that one of the respondents on the multiple choice also provided a shorter, less precise answer in question 1 ('phenotypes' instead of 'loss of function phenotypes'), even though both were given as options on the questionnaire.

Further research is needed to determine if the multiple choice format is more reliable when there is a larger number and range of examples and how to assemble the possible choices in a more principled way.

Acknowledgements

The authors would like to acknowledge the support of the University of Minnesota Graduate School (Initiatives in Interdisciplinary Research, Scholarly and Creative Activities and Grant-in-Aid Programs) and the Digital Technology Center. This work was also partially supported by NSF Grant IIS-0208621.

References

Butler, B.; Sproull, L.; Kiesler, S.; and Kraut, R. 2005. Community building in online communities, who does the work and why? In Weisband, S., and Atwater, L., eds., *Leadership at a Distance*. Erlbaum. forthcoming.

Castaño, J.; Zhang, J.; and Pustejovsky, J. 2002. Anaphora resolution in biomedical literature. In *International Symposium on Reference Resolution*.

Chklovski, T., and Mihalcea, R. 2003. Exploiting agreement and disagreement of human annotators for word sense disambiguation. In *Proc. of the Conference on Recent Advances in Natural Language Processing (RANLP)*.

Gundel, J.; Hedberg, N.; and Zacharski, R. 1993. Cognitive status and the form of referring expressions in discourse. *Language* 69:274–307.

Hobbs, J. 1978. Resolving pronoun references. *Lingua* 44:311–338.

Lappin, S., and Leass, H. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics* 20(4):535–561.

Mihalcea, R., and Chklovski, T. 2004. Building sense tagged corpora with volunteer contributions over the web. In Nicolov, N., and Mitkov, R., eds., *Current Issues in Linguistic Theory: Recent Advances in Natural Language Processing*. John Benjamins Publishers.

Mitkov, R. 2002. *Anaphora Resolution*. Longman.

Ng, V., and Cardie, C. 2002. Improving machine learning approaches to coreference resolution. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 104–111.

Popescu-Belis, A.; Rigouste, L.; Salmon-Alt, S.; and Romary, L. 2004. Online evaluation of coreference resolution. In *Proc. of the 4th Int'l Conference on Language Resources and Evaluation (LREC)*.

Preece, J. 2000. *Online Communities: Designing Usability and Supporting Sociability*. John Wiley.

Salmon-Alt, S., and Romary, L. 2004. Data categories for a normalized reference annotation scheme. In *Proc. of the 5th Discourse, Anaphor Resolution Colloquium (DARRC)*, 145–150.

Schütze, C. T. 1996. *The Empirical Base of Linguistics. Grammaticality Judgments and Linguistic Methodology*. University of Chicago Press.

Soon, W. M.; Ng, H. T.; and Lim, D. C. Y. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics* 27(4):521–544.

Strube, M., and Müller, C. 2003. A machine learning approach to pronoun resolution in spoken dialogue. In *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Yang, X.; Zhou, G.; Su, J.; and Tan, C. L. 2003. Coreference resolution using competitive learning approach. In *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, 176–183.