ANAWIKI: Creating Anaphorically Annotated Resources Through Web Cooperation

Massimo Poesio[♠], Udo Kruschwitz[♣], Jon Chamberlain[♣]

[♠]Uni Essex and Uni Trento
[♣]Uni Essex

Abstract

The ability to make progress in Computational Linguistics depends on the availability of large annotated corpora, but creating such corpora by hand annotation is very expensive and time consuming; in practice, it is unfeasible to think of annotating more than one million words. However, the success of Wikipedia, the ESP game, and other projects shows that another approach might be possible: collaborative resource creation through the voluntary participation of thousands of Web users. ANAWIKI is a recently started project that will develop tools to allow and encourage large numbers of volunteers over the Web to collaborate in the creation of annotated corpora (in the first instance, of a corpus annotated with semantic information about anaphoric relations) through a variety of interfaces.

1. Introduction

Perhaps the greatest obstacle to the development of systems able to extract semantic information from text is the lack of semantically annotated corpora large enough to train and evaluate semantic interpretation methods; however, the creation of semantically annotated corpora has lagged behind the creation of corpora for other types of NLP subtasks. Recent efforts in the USA to create resources to support semantic evaluation initiatives such as Automatic Context Extraction (ACE), Translingual Information Detection, Extraction and Summarization (TIDES), and Global Autonomous Language Exploitation GALE are beginning to change this situation through the creation of 1 Million word annotated corpora such as PropBank (Palmer et al., 2005) and the OntoNotes initiative (Hovy et al., 2006), but just at a time when the community begins to realize that even such corpora are too small. Unfortunately, the creation of 100M-plus corpora via hand-annotation is likely to be prohibitively expensive, as already realized by the creators of the British National Corpus (Burnard, 2000), much of whose annotation was done automatically. Such a large hand-annotation effort would be even less sensible in the case of semantic annotation tasks such as coreference or wordsense disambiguation, given, on one side, the greater difficulty of agreeing on a 'neutral' theoretical framework for the annotation, an essential prerequisite (see, e.g., discussion in (**?**; Palmer et al., 2005)); on the other, the difficulty of achieving more than moderate agreement on semantic judgments (Poesio and Artstein, 2005; Zaenen, 2006). For this reason, a great deal of effort is underway to develop and/or improve semi-automatic methods for creating annotated resources and/or for using the existing data, such as active learning and bootstrapping.

The primary objective of the ANAWIKI project[1] is to evaluate an as yet untried approach to the creation of large-scale annotated corpora: taking advantage of the collaboration of the Web community.

2. Background

2.1. Alternatives to hand-annotation

The best-known approach to overcome the limitations of hand annotation is to combine a small, high-quality hand annotated corpus with additional unannotated monolingual text as input to a bootstrapping process that attempts to extend annotation coverage to the corpus as a whole. A number of techniques for combining labelled and unlabelled data exist. CO-TRAINING METHODS (Blum and Mitchell, 1998; Yarowsky, 1995) make crucial usage of redundant models of the data. A family of approaches known as ACTIVE LEARNING have also been developed, in which some of the unlabelled data are selected and manually labelled (e.g., (Vlachos, 2006)). NLP systems trained on this labelled data almost invariably produce better results than systems trained on data which was not specially selected and certainly better than data which was labelled using automatic means such as EM or co-training. Finally, in the case of parallel corpora where one of the languages has already been annotated, projection techniques (Diab and Resnik, 2002) can be used to 'transfer' the annotation to other languages. None of these techniques however produces improvements comparable to those achievable with more annotated data.

2.2. Creating Resources for AI through Web collaboration

To our knowledge, ours is the first attempt to exploit the effort of Web volunteers to create annotated corpora; but there have been other attempts at harnessing the Web community to create knowledge. Wikipedia is perhaps the best example of collective resource creation, but it is not an isolated case. The willingness of Web surfers to help extends to projects to create resources for Artificial Intelligence. One example is the Open Mind Commonsense project (Singh, 2002),[2] a project to mine commonsense knowledge to which 14500 participants contributed nearly 700,000 facts. If the 15,000 volunteers who participated in Open Mind Commonsense each annotated 7 texts of 1,000 words (an effort of about 3 hours), we would get a 100M

[1] www.anawiki.org/

[2] commonsense.media.mit.edu/

words annotated corpus. A more recent, and even better known, example is the ESP game,[3] a project to label images with tags through a competitive game (von Ahn, 2006).

A slightly different approach to the creation of commonsense knowledge has been pursued in the Semantic Wiki project,[4] an effort to develop a 'Wikipedia way to the Semantic Web': i.e., to make Wikipedia more useful and to support improved search of web pages via semantic annotation. The project is implementation-oriented, with a focus on the development of tools for the annotation of concept instances and their attributes, using RDF to encode such relations. Software (still in beta version) has been developed; apart from the Semantic MediaWiki software itself (which builds on MediaWiki, the software running Wikipedia), tools independent from Wikipedia have also been developed, including IkeWiki, OntoWiki, Semantic Wikipedia, etc.

3. Methods

3.1. Annotating anaphoric information

ANAWIKI will build on the proposals for marking anaphoric information allowing for ambiguity developed in ARRAU (Poesio and Artstein, 2005) and previous projects (Poesio, 2004). In these projects we developed, first of all, instructions for volunteers to mark anaphoric information in a reliable way that were tested in a series of agreement studies. These instructions specify how to annotate the fundamental types of anaphoric information, including both identity relations and the most basic types of bridging relations, and how to mark different types of ambiguity. Crucially, these instructions were tested in a series of experiments in which annotators received very little training, which is very close to the way annotation will be carried out with ANAWIKI. Second, we developed XML markup standards for coreference that make it possible to encode alternative options. Also relevant for the intended work are the findings from ARRAU that (i) using numerous annotators (up to 20 in some experiments) leads to a much more robust identification of the major interpretation alternatives (although outliers are also frequent); and (ii) the identification of alternative interpretations is much more frequently a case of implicit ambiguity (each volunteer identifies only one interpretation, but these are different) than of explicit ambiguity (volunteers identifying multiple interpretations).

In ARRAU we also developed methods to analyze collections of such alternative interpretations and to identify outliers via clustering that we will exploit in the proposed project. These methods for representing multiple interpretations and for dealing with them will be used in the proposed project as the technical foundation for an annotation tool making it possible for multiple Web volunteers to annotate semantic information in a text without losing previous information.

3.2. Expert and Game Interfaces

The objective of ANAWIKI is to create tools that will make it possible (in fact, interesting) for volunteers to annotate anaphoric information in texts over the Web. It is expected that different types of volunteers will prefer different types of interface, and we plan at least two. A more conventional interface of the type familiar from annotation tools will be developed for (computational) linguists who want to cooperate in the effort of creating such resources on the Web, through the University of Bielefeld's Serengeti software.[5] In addition, we are developing and testing a game-like interface as in the case of ESP game, which may appeal to a larger group of Web volunteers.

3.3. Monotonic annotation

Both Wikipedia and Semantic Wikis are based on the same assumption: that there is only one version of the text, and subsequent editors dissatisfied with previous versions of the text replace portions of it (under some form of supervision). A version history is preserved, but it is only accessible through diff files. But work on semantic annotation in Computational Linguistics, and particularly the (mediocre) results at evaluating agreement among annotators, made it abundantly clear that for many types of semantic information in text–wordsense above all (Véronis, 1998) but also coreference (Poesio and Artstein, 2005) –small differences in interpretation are the rule and often major differences are possible, especially when fine-grained meaning distinctions are required (Poesio and Artstein, 2005). Forcing volunteers to agree on a single interpretation is likely to lead to dispute, just as in the case of disagreements on the content of pages in Wikipedia, and it precludes collecting data on disagreements which will allow us to understand which semantic judgments, if any, lead to fewer disagreements. In ANAWIKI, we will adopt a monotonic approach to annotation, in which all interpretations which pass certain tests will be maintained.

The expert volunteers will be able to see all annotations of that markable by other volunteers (all semantic relations in which that markable enters) and express either agreement or disagreement with these annotations. They will also be able to add new annotations, by clicking on another markable and then specifying the relation between them (one of a fixed range). Our experience in ARRAU suggests that allowing for a maximum of 10 alternative interpretations, a number which could be visualized with a menu, should be adequate. The participants in the game will be able to 'validate' the proposals of other gamers.

3.4. Instructions for volunteers

There is always a tension in annotation projects between the goals of collecting actual judgments and that of obtaining interpretable results. In the case of AnaWiki there are further constraints, as experts may already know about the task (although their views may be different from ours) whereas gamers will quickly get bored if we impose too many constraints on what they can do. On the other end, in order for the results to be usable for the purposes of semantic interpretation, it will be necessary for the volunteers to understand the type of semantic judgments we are interested in; it

[3] www.espgame.org/

[4] www.semwiki.org/

[5] coli.lili.uni-bielefeld.de/serengeti/annotator.pl

will also be necessary to find ways of identifying malicious or simply lazy volunteers.

As mentioned above, ARRAU produced a detailed manual for the annotation of coreference; this manual will be adapted to use with the expert interface in ANAWIKI, tested in a small study at Essex, and put on the website for volunteers to read. We will test how well each volunteer understands the instructions by asking them to mark a simple text to start with and analyzing what they do using outlier detection methods also developed in ARRAU. (If the results seem to indicate poor understanding, they will still be able to annotate, but the results will be discarded.) A novel aspect of the proposed project is the fact that volunteers will also be able to comment on other volunteers' output, which hopefully will give us another weapon for identifying outliers.

For the gamers, we are experimenting with giving the instructions through a training process in which they get more points the more they act in agreement with the instructions.

3.5. Markable Identification and Editing

A familiar interface–e.g., a Wiki-style interface– may reduce training time, but may also introduce problems, in that an interface for annotation purposes differs from one used for information creation in a number of ways. A first difference is text modification: users of the ANAWIKI tools will not be able to modify the texts they are annotating. A second difference is markable identification.

One of the guiding criteria for our interface design will be to make the task as easy and fun as possible; but it would be a lot of work for the volunteers to also identify all parts of text that identify objects that enter into semantic relations (markables). Instead, each such text constituent will be automatically pre-marked and assigned a unique index[6] so that volunteers will only be required to click inside a markable in order to be able to specify semantic relations with that markable as first argument. We will however experiment with allowing our collaborators to correct the errors of the parser by, e.g., adding / removing markables or changing markable boundaries.

3.6. The data

One of the biggest problems with current semantically annotated corpora (unlike, say, the BNC) is that they are not balanced–in fact, they tend to consist almost exclusively of news articles. We plan to address this issue by including a selection of English texts from different domains and different genres. We think this will also make the task more interesting for the volunteers. Only copyright-free texts will be included.

One obvious example of texts not extensively represented in current semantically annotated corpora, yet central to the study of language, is narratives. Fortunately, a great deal of narrative text is available copyright-free, e.g., through Project Gutenberg for English and similar initiatives for other languages.

Another example of texts not included in current semantically annotated corpora are encyclopaedic entries like those from Wikipedia itself: they are available in both English and Italian, are the main target for the Semantic Media Wiki project, and have already been created using this software. We also expect to include sample text from emails (e.g., from the Enron corpus) and transcripts of spoken text.

3.7. The Anaphoric Bank

ANAWIKI is part of an effort to overcome difficulties in securing funding for annotation efforts, particularly in Europe, by create the Anaphoric Bank[7]–a multi-lingual anaphorically annotated corpus–through the collaboration of a number of sites each of which has been carrying out a limited amount of annotation. The Anaphoric Bank will be organized as a 'club' which can be entered by paying a 'fee' of a certain amount of data annotated using one of the compatible formats and annotation schemes; membership will then give access to the rest of the annotated resources. [8]

From a technical point of view, the Anaphoric Bank is made possibile by the development of so-called 'pivot' standoff XML formats such as Paula,[9] into which data annotated with one of the several existing tools (MMAX, PALINKA, etc) can be imported and subsequently exported.

3.8. WebSite

All of the different interfaces will be accessible through a web server, that will also be used to coordinate the annotation activities for experts.

The website with the expert interface will contain (i) the instructions describing the task; (ii) a way to register as volunteers; (iii) simple tests that volunteers must 'pass' in order for their annotation to be included in the standard (see below) (iv) an indication of which texts still need work (need to be annotated or reannotated), as with Wikipedia; (v) methods for allowing the contributors to the AnaphoricBank project to export the annotated data in XML format. We also believe it is key for the website to be linked to international sites. We plan to reach agreements in this respect with the Semantic MediaWiki site, as well as the relevant professional societies–e.g., AAAI, ACH, ACL, ACM, the Cognitive Science Society, and IATH.

For the game interface, we have developed a website which emphasizes the gaming aspects and in which the instructions are targeted towards playing the game. This site, too, will be linked to international sites, but we will particularly target sites such as Flickr, etc.

4. Conclusions

We presented the ANAWIKI project, an effort to develop collaborative methods of annotation which started at the University of Essex in November 2007 and will run to May 2009. We would welcome collaboration with other groups

[6] We will use for this purpose a customized version of the NLP pipeline of the BART system–see (**?**), this conference.

[7] www.anaphoricbank.org

[8] Current members of the consortium include University of Bielefeld, EML Research, University of Essex, University of Potsdam, UNISINOS, University of Trento, and University of Wolverhampton.

[9] www.sfb632.uni-potsdam.de/homes/d1/paula/doc/

on this effort to find new methods for resource creation for Computational Linguistics.

Acknowledgments

ANAWIKI is funded EPSRC, grant EP/F00575X/1. We wish to thank all the participants in the Anaphoric Bank consortium, and in particular Daniela Goecke, Nils Diewald, Maik Stührenberg, and Dieter Metzing from Uni Bielefeld. We also wish to thank Ron Artstein for comments and his contributions to the ARRAU project.

5. References

Blum and Mitchell. 1998. Combining labelled and unlabelled data with co-training. In *Proc. 11th Conference on Computational Learning Theory,*, pages 92–100.

L. Burnard. 2000. The British National Corpus reference guide. Technical report, Oxford University Computing Services, Oxford. `http://www.natcorp.ox.ac.uk/World/HTML/urg.html`.

M. Diab and P. Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proc. of ACL 2002*, Philadelphia.

E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. 2006. Ontonotes: the 90% solution. In *Proc. HLT-NAACL*.

M. Palmer, D. Gildea, and Kingsbury. 2005. The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics*.

M. Poesio and R. Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In A. Meyers, editor, *Proc. of ACL Workshop on Frontiers in Corpus Annotation*, pages 76–83, June.

M. Poesio. 2004. The MATE/GNOME scheme for anaphoric annotation, revisited. In *Proc. of SIGDIAL*, Boston, May.

P. Singh. 2002. The public acquisition of commonsense knowledge. In *Proc. AAAI Spring Symposium on Acquiring (and using) Linguistic (and World) knowledge for information access*.

J. Véronis. 1998. A study of polysemy judgments and inter-annotator agreement. In *Proc. of SENSEVAL-1*.

A. Vlachos. 2006. Active annotation. In *Proc. EACL 2006 Workshop on Adaptive Text Extraction and Mining*, Trento.

Luis von Ahn. 2006. Games with a purpose. *Computer*, 39(6):92–94.

D. Yarowsky. 1995. Unsupervised wordsense disambiguation rivalling supervised methods. In *Proc. 33rd ACL*.

A. Zaenen. 2006. Mark-up barking up the wrong tree. *Computational Linguistics*, 32(4):577–580.