(Linguistic) Science Through Web Collaboration in the ANAWIKI Project

Udo Kruschwitz
University of Essex
udo@essex.ac.uk

Jon Chamberlain
University of Essex
jchamb@essex.ac.uk

Massimo Poesio
University of Essex and
Università di Trento
poesio@essex.ac.uk

Abstract

Despite the impressive progress made in recent years in all areas of natural language processing there are still tasks that do not perform well enough to be used in everyday applications. One example is anaphora resolution. The most promising approach to get significant improvements in this area is to create sufficiently large linguistically annotated resources which can then be used to train, for example, machine learning systems. Annotated corpora of the size needed for modern computational linguistics research cannot however be created by small groups of hand-annotators; but ESP and similar games have demonstrated how it might be possible to do this through Web collaboration. This paper reports on the ongoing work on *Phrase Detectives*, a game developed in the ANAWIKI project designed for collaborative linguistic annotation on the Web. Of particular concern here are the measures that assure high-quality annotations.

1 Introduction

The statistical revolution in natural language processing (NLP) has resulted in the first NLP systems and components really usable on a large scale, from part-of-speech (POS) taggers to parsers [7]. But it has also raised the problem of creating the large amounts of annotated linguistic data needed for training and evaluating such systems. Potential solutions to this problem include semi-automatic annotation, and machine learning methods that make better use of the available data. Unsupervised or semi-supervised techniques hold great promise, but for the foreseeable future at least, the greatest performance improvements are still likely to come from increasing the amount of data to be used by supervised training methods, which crucially rely on hand-annotated data. Traditionally, this requires trained annotators, which is prohibitively expensive both financially and in terms of person-hours (given the number of trained annotators available) on the scale required.

Recently, however, web collaboration has started to emerge as a viable alternative. Wikipedia and similar initiatives have shown that a surprising number of individuals are willing to help with resource creation and scientific experiments. The Open Mind Common Sense project [12] demonstrated that such individuals are also willing to participate in the creation of databases for Artificial Intelligence (AI), and von Ahn showed that web games are an effective way of motivating subjects to annotate data for machine learning purposes [16, 17].

The goal of the ANAWIKI project[1] is to experiment with Web collaboration as a solution to the problem of creating large-scale linguistically annotated corpora, both by developing tools through which members of our scientific community can participate in corpus creation through annotation tools with a Web interface and through the use of game-like interfaces. We will present ongoing work on *Phrase Detectives*[2], a game designed to collect judgments about anaphoric annotations. We will also report results which include a substantial corpus of annotations that has already been collected.

The paper will be structured as follows. We will start with a brief discussion of some related work (Section 2) and address the main issues arising (Section 3). We will then introduce the *Phrase Detectives* game (Section 4) followed by a discussion of how we enforce quality control (Section 5). Two further sections will discuss implementational issues and first results before we outline future work.

[1] http://www.anawiki.org
[2] http://www.phrasedetectives.org

2 Related Work

Related work comes from a range of fairly distinct research communities including, among others, Computational Linguistics / NLP as well as the games community, but also from researchers working in the areas of the Semantic Web and knowledge representation.

Large-scale annotation of low-level linguistic information (part-of-speech tags) began with the Brown Corpus, in which very low-tech and time consuming methods were used; but already for the creation of the British National Corpus (BNC), the first 100M-word linguistically annotated corpus, a faster methodology was developed consisting of preliminary annotation with automatic methods followed by partial hand-correction [2]. This was made possible by the availability of fairly high-quality automatic part-of-speech taggers (CLAWS). With the development of the first medium high-quality chunkers this methodology became applicable to the case of syntactic annotation, and indeed was used for the creation of the Penn Treebank [8] although in this case much more substantial hand-checking was required.

Medium and large-scale semantic annotation projects (coreference, wordsense) are a fairly recent innovation in Computational Linguistics. The semi-automatic annotation methodology cannot yet be used for this type of annotation, as the quality of, for instance, coreference resolvers is not yet high enough on general text. Nevertheless the semantic annotation methodology has made great progress with the development, on the one end, of effective quality control methods (see for example [4]); on the other, of sophisticated annotation tools such as Serengeti [15]. These developments have made it possible to move from the small-scale semantic annotation projects of a few years ago, whose aim was to create resources of around 100K words in size, e.g. [10], to the efforts made as part of US initiatives such as Automatic Context Extraction (ACE), Translingual Information Detection, Extraction and Summarization (TIDES), and GALE to create 1M words corpora. But such techniques could not be expected to be used to annotate data on the scale of the BNC.

Collective resource creation on the Web offers a different way to the solution of this problem. The motivation for this is the observation that a group of individuals can contribute to a collective solution which has a better performance and is more robust than an average individual's solution as demonstrated in simulations of collective behaviours in self-organizing systems [6].

Wikipedia is perhaps the best example of collective resource creation, but it is not an isolated case. The gaming approach to data collection, termed *Games with a purpose*, has received increased attention since the success of the ESP game [16]. Subsequent games have attempted to collect data for multimedia tagging (*OntoTube*, *Tag a Tune*) and language tagging (*Verbosity, OntoGame, Scattergories, Taboo*). As Wikipedia has demonstrated however, there is not necessarily the need to turn every data collection task into a game. Other current efforts in attempting to acquire large-scale world knowledge from Web users include Freebase[3] and True Knowledge[4].

Interestingly, the *Games with a purpose* concept has now also been adopted by the Semantic Web community in an attempt to collect large-scale ontological knowledge because currently "the Semantic Web lacks sufficient user involvement almost everywhere" [13].

3 Issues Arising

In order to use Web collaboration to create annotated data, a number of serious issues have to be addressed. First among these is motivation. For anybody other than a few truly dedicated people, annotation is a very boring task. This is where the promise of the game approach lies. Provided that a suitably entertaining format can be found, it may be possible to get people to tag quite a lot of data without them even realizing it. (Of course, finding such a format is by no means trivial.)

Assuming that this can be done, other problems still remain, most important of which is to ensure the *quality* of the annotated data. We have identified the following four aspects that need to be addressed to control annotation quality:

- Ensuring users understand the task
- Attention slips
- Malicious behaviour
- Genuine ambiguity of data

We will discuss how each of these points has been addressed in *Phrase Detectives*. First however, we will give an overview of the game.

[3] `http://www.freebase.com/`
[4] `http://www.trueknowledge.com/`

4 The Phrase Detectives Game

Phrase Detectives is a game offering a simple interface for non-expert users to learn how to annotate text and to make annotation decisions. The goal of the game is to identify relationships between words and phrases in a short text. An example of a task would be to highlight an anaphor-antecedent relation between the "markables" (sections of text) *'This parrot'* and *'He'* in *'This parrot is no more! He has ceased to be!'* Markables are identified in the text by automatic pre-processing. There are two ways to annotate within the game: by selecting a markable that corefers to another one (Annotation Mode); or by validating a decision previously submitted by another player (Validation Mode).

Annotation Mode is the simplest way of collecting judgments. The player has to locate the closest antecedent markable of an anaphor markable, i.e. an earlier mention of the object. By moving the cursor over the text, markables are revealed in a bordered box. To select it the player clicks on the bordered box and the markable becomes highlighted. They can repeat this process if there is more than one antecedent markable (i.e. for plural anaphors such as "they"). They submit the annotation by clicking the "Found it!" button and are given points. The player can also, among other options, indicate that the highlighted markable has not been mentioned before (i.e. it is not anaphoric), or they can skip the markable and move on to the next one.

In Validation Mode (see Figure 1) the player is presented with an annotation from a previous player. The anaphor markable is shown with the antecedent markable(s) that the previous player chose. The current player has to decide if they agree with this annotation. Points are given to the current player, and also to the previous player who made the original annotation. If the current player disagrees with the previous one he is shown the Annotation Mode so he can enter a new annotation.

5 Quality Control

We will now look at how we address the four main quality concerns.

5.1 Task Comprehension

The first ingredient of the *Phrase Detectives* approach to quality control is to use a training level that ensures only players that display an understanding of the rules of the game get to annotate the data of interest. Players begin the game at the training level where they are given a set of annotation tasks created from the Gold Standard. They are given feedback and guidance when they select an incorrect answer and points when they select the correct answer. When the player gives enough correct answers they graduate to real annotation; it is only the annotations of these players that are included in the corpus.

From time to time, a graduated player will be covertly given a Gold Standard text to annotate. A bonus screen is shown when the player has completed annotating the text indicating what the player selected incorrectly, with bonus points for agreeing with the Gold Standard. This is the foundation of a player rating system to judge the quality of the player's annotations.

The game is designed to motivate players to annotate the text correctly by using comparative scoring (e.g., awarding points for agreeing with the Gold Standard), and collaborative scoring (awarding points to the current player if they are agreed with by the previous player). Using leader boards and assigning levels for points has been proven to be an effective motivator, with players often using these as targets [17].

5.2 Attention Slips

Players may occasionally make a mistake and press the wrong button. We have made a deliberate decision that there is no way that a player could go back and try again (e.g. imagine a player trying out all possible annotations and then selecting the one offering the highest score). How do we address this problem that we could call "attention slip"? Primarily by applying the Validation Mode, our second strategy for quality control, where players can examine other players' annotations and evaluate them. Validation Mode plays three essential roles in the game: addressing attention slips, filtering dubious annotations (i.e., letting the players serve a function similar to that of judges in Wikipedia) and acting as a social training mechanism where players can see what other players have done.

5.3 Malicious Input

Several methods are used to identify players who are cheating or who are providing poor annotations. These include checking the player's IP address, checking annotations against known answers and keeping a blacklist of players to discard all their data [16]. The prime method however to filter out malicious input is through validation. In

Figure 1: A screenshot of the Validation Mode.

the same way, as, for example, the PageRank algorithm imposes a relevance order on Web pages based on the Web's link structure [1], we assess the quality of annotations by the agreement level between players (to stick with the analogy, validations can be seen as incoming links and player ratings can be authority values of each validated link). All annotations are validated by a number of different players, essentially filtering out malicious input.

5.4 Genuine Ambiguity

Ambiguity is an inherent problem in all areas of NLP [7]. Here we are not interested in solving this issue but in capturing ambiguity where it is appropriate. If an anaphor is ambiguous, then the annotated corpus should capture this information. We are therefore not aiming at selecting "the best" or most common annotation but to preserve all inherent ambiguity (which is supported by the output formats discussed below).

6 Implementation

Phrase Detectives is running on a dedicated Linux server. The pre-processed data is stored in an MySQL database, most of the scripting is done via PHP. The game is not however an isolated stand-alone implementation. One aspect of the project not discussed in this paper is an expert annotation tool (closely linked to *Phrase Detectives*) which we use to obtain Gold Standard annotations. The Gold Standard is used to train the players and to check their annotation quality in regular intervals. The expert annotation tool is used by computational linguists. In the case of anaphora annotation we use the Serengeti tool developed at the University of Bielefeld [15]. This tool runs on the same server and accesses the same database.

The database stores the textual data in Sekimo Generic Format (SGF) [14], a multi-layer representation of the original documents that can easily be transformed into other common formats such as MAS-XML and PAULA. We apply a pipeline of scripts to get from raw text to SGF format. For English texts this pipeline consists of these main steps:

- A pre-processing step normalises the input, applies a sentence splitter and runs a tokenizer over each sentence. We use the *openNLP*[5] toolkit to perform this process.
- Each sentence is then analysed by the *Berkeley Parser*[6].
- The parser output is interpreted to identify markables in the sentence. As a result we create an XML representation which preserves the syntactic structure of the markables (including nested markables, e.g. noun phrases within a larger noun phrase).
- A heuristic processor identifies a number of additional features associated with markables such as person, case, number etc. The output format is MAS-XML.

[5] `http://opennlp.sourceforge.net/`
[6] `http://nlp.cs.berkeley.edu/`

The last two steps are based on previous work within the research group [11]. Finally, MAS-XML is converted into SGF. Both MAS-XML and SGF are also the formats used to export the annotated data.

7 Results

Before going live we evaluated a prototype of the game interface informally by a group of randomly selected volunteers from the University of Essex [3]. The beta version of *Phrase Detectives* went on-line in May 2008, with the first live release in December 2008. Initially over 100,000 words of text from Project Gutenberg and Wikipedia were automatically parsed to identify the markables and added to the game. The game now accesses a corpus of more than a million words.

In less than three months of live release the game has collected over 100,000 annotations of anaphoric relations provided by more than 500 players. To put this in perspective, the GNOME corpus, produced by traditional methods, included around 3,000 annotations of anaphoric relations [9] whereas OntoNotes[7] 3.0, with 1 million words, contains around 140,000 annotations.

The analysis of the results is an ongoing issue. However, by manually analyzing 10 random documents we could not find a single case in which a misconceived annotation was validated by other players. This confirms the assumptions we made about quality control but will obviously need to be further investigated by more thorough analysis methods which will be part of the future work.

8 Future Work

We are progressively converting text for use in the game with the aim of having 100 million words. So far, mainly narrative texts from Project Gutenberg and encyclopedic texts from Wikipedia have been converted; we also plan to include further data from travel guides, news articles, and the American National Corpus [5].

We will make the annotated data available to the community through the Anaphoric Bank[8].

Ultimately, the usefulness of the annotated data will need to be shown by, for example, successfully training an anaphora resolution algorithm that performs better than existing systems.

[7] `http://www.ldc.upenn.edu/`

[8] `http://www.anaphoricbank.org/`

Acknowledgements

ANAWIKI is funded by a grant from the Engineering and Physical Sciences Research Council (EPSRC), grant number EP/F00575X/1. Thanks to Daniela Goecke, Maik Stührenberg, Nils Diewald and Daniel Jettka at the University of Bielefeld who have been closely involved in the development of this project. We also want to thank all volunteers who have already contributed to the project, in particular our current top players *trelex* and *livio.robaldo*.

References

[1] Brin, S., and Page, L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proceedings of the Seventh International World Wide Web Conference (WWW7)* (Brisbane, 1998), pp. 107–117.

[2] Burnard, L. The British National Corpus Reference guide. Technical report, Oxford University Computing Services, Oxford, 2000.

[3] Chamberlain, J., Poesio, M., and Kruschwitz, U. Phrase Detectives: A Web-based Collaborative Annotation Game. In *Proceedings of the International Conference on Semantic Systems (I-Semantics'08)* (Graz, 2008).

[4] Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. OntoNotes: The 90% Solution. In *Proceedings of HLT-NAACL06* (2006).

[5] Ide, N., and Macleod, C. The American National Corpus: A Standardized Resource of American English. In *Proceedings of Corpus Linguistics* (Lancaster, 2001).

[6] Johnson, N. L., Rasmussen, S., Joslyn, C., Rocha, L., Smith, S., and Kantor, M. Symbiotic Intelligence: Self-Organizing Knowledge on Distributed Networks Driven by Human Interaction. In *Proceedings of the Sixth International Conference on Artificial Life* (1998), MIT Press.

[7] Jurafsky, D., and Martin, J. H. *Speech and Language Processing- 2^{nd} edition.* Prentice-Hall, 2008.

[8] Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. Building a large annotated corpus of english: the Penn Treebank. *Computational Linguistics 19*, 2 (1993), 313–330.

[9] Poesio, M. Discourse annotation and semantic annotation in the gnome corpus. In *Proceedings of the ACL Workshop on Discourse Annotation* (2004).

[10] Poesio, M. The MATE/GNOME scheme for anaphoric annotation, revisited. In *Proceedings of SIGDIAL* (2004).

[11] Poesio, M., and Artstein, R. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation* (2005), pp. 76–83.

[12] Singh, P. The public acquisition of commonsense knowledge. In *Proceedings of the AAAI Spring Symposium on Acquiring (and Using) Linguistic (and World) Knowledge for Information Access* (Palo Alto, CA, 2002).

[13] Siorpaes, K., and Hepp, M. Games with a purpose for the semantic web. *IEEE Intelligent Systems 23*, 3 (2008), 50–60.

[14] Stührenberg, M., and Goecke, D. SGF - An integrated model for multiple annotations and its application in a linguistic domain. In *Proceedings of Balisage: The Markup Conference* (Montreal, 2008).

[15] Stührenberg, M., Goecke, D., Diewald, N., Mehler, A., and Cramer, I. Web-based annotation of anaphoric relations and lexical chains. In *Proceedings of the ACL Linguistic Annotation Workshop* (2007), pp. 140–147.

[16] von Ahn, L. Games with a purpose. *Computer 39*, 6 (2006), 92–94.

[17] von Ahn, L., and Dabbish, L. Designing games with a purpose. *Communications of the ACM 51*, 8 (2008), 58–67.