Proceedings of I-SEMANTICS '08
Graz, Austria, September 3-5, 2008

Phrase Detectives: A Web-based Collaborative Annotation Game

Jon Chamberlain
(University of Essex, Colchester, UK
jchamb@essex.ac.uk)

Massimo Poesio
(University of Essex, Colchester, UK and Università di Trento, Trento, Italy
poesio@essex.ac.uk)

Udo Kruschwitz
(University of Essex, Colchester, UK
udo@essex.ac.uk)

Abstract: Annotated corpora of the size needed for modern computational linguistics research cannot be created by small groups of hand annotators. One solution is to exploit collaborative work on the Web and one way to do this is through games like the ESP game. Applying this methodology however requires developing methods for teaching subjects the rules of the game and evaluating their contribution while maintaining the game entertainment. In addition, applying this method to linguistic annotation tasks like anaphoric annotation requires developing methods for presenting text and identifying the components of the text that need to be annotated. In this paper we present the first version of *Phrase Detectives* (`http://www.phrasedetectives.org`), to our knowledge the first game designed for collaborative linguistic annotation on the Web.

Key Words: Web-based games, distributed knowledge acquisition, object recognition, social networking, anaphoric annotation, user interaction, XML, Semantic Web
Category: H.5.2, I.2.5, I.2.6, I.2.7

1 Introduction

Perhaps the greatest obstacle to progress towards systems able to extract semantic information from text is the lack of semantically annotated corpora large enough to be used to train and evaluate semantic interpretation methods. Recent efforts to create resources to support large evaluation initiatives in the USA such as Automatic Context Extraction (ACE), Translingual Information Detection, Extraction and Summarization (TIDES), and GALE are beginning to change this – but just at a point when the community is beginning to realize that even the 1M word annotated corpora created in substantial efforts such as Prop-Bank [Palmer et al., 2005] and the OntoNotes initiative [Hovy et al., 2006] are likely to be 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 the one side the greater difficulty of agreeing on a 'neutral' theoretical framework, 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 (`http://www.anawiki.org`) is to experiment with a novel approach to the creation of large-scale annotated corpora: taking advantage of the collaboration of the Web community, both through co-operative annotation efforts using traditional annotation tools and through the use of game-like interfaces [Poesio et al., 2008]. In this paper we present our work to develop *Phrase Detectives*, a game designed to collect judgments about anaphoric annotations.

2 Creating Resources

2.1 Traditional Annotation Methodology

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 [Burnard, 2000]. 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 [Marcus et al., 1993] 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 semantic annotation methodology has made great progress with the development, on the one end, of effective quality control methods (see for example [Hovy et al., 2006]); on the other, of sophisticated annotation tools such as Serengeti [Stührenberg et al., 2007]. 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. [Poesio, 2004], to projects aiming at creating 1M words corpora. But such techniques could not be expected to be used to annotate data on the scale of the British National Corpus.

2.2 Creating Resources through Web Collaboration

Collective resource creation on the Web offers a different way to the solution of this problem. Wikipedia is perhaps the best example of collective resource creation, but it is not an isolated case. The willingness of Web users to volunteer on the Web extends to projects to create resources for Artificial Intelligence. One example is the Open Mind Commonsense project, a project to mine commonsense knowledge to which 14,500 participants contributed nearly 700,000 sentences [Singh, 2002]. Current efforts in attempting to acquire large-scale world knowledge from Web users include Freebase (`http://www.freebase.com/`) and True Knowledge (`http://www.trueknowledge.com/`).

A slightly different approach to the creation of commonsense knowledge has been pursued in the Semantic MediaWiki project [Krötzsch et al., 2007], 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.

A perhaps more intriguing development is the use of interactive game-style interfaces to collect knowledge such as LEARNER [Chklovski and Gil, 2005], Phetch, Verbosity and Peekaboom [von Ahn et al., 2006]. The ESP game is perhaps the best known example of this approach, a project to label images with tags through a competitive game. 13,500 users played the game, creating 1.3M labels in 3 months [von Ahn, 2006]. If we managed to attract 15,000 volunteers, and each of them were to annotate 10 texts of 700 words, we would get a corpus of the size of the BNC.

2.3 Annotating Anaphoric Information

ANAWIKI builds on the proposals for marking anaphoric information allowing for ambiguity developed in ARRAU [Poesio and Artstein, 2005] and previous projects [Poesio, 2004]. The ARRAU project found 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 annotator identifies only one interpretation, but these are different) than of explicit ambiguity (annotators identifying multiple interpretations). The ARRAU project also developed methods to analyze collections of such alternative interpretations and to identify outliers via clustering that will be exploited in this project. These methods for representing multiple

interpretations and for dealing with them are used as the technical foundation for an annotation tool making it possible for multiple Web volunteers to annotate semantic information in text.

3 Game Interface for Annotating Data

3.1 Description of the Game

Phrase Detectives is a game offering a simple user 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. "Markables" are identified in the text by automatic pre-processing. There are 2 ways to annotate within the game: by selecting a markable that corefers to another highlighted markable (Annotation Mode - see Figure 1); or by validating a decision previously submitted by another user (Validation Mode - see Figure 2).

3.2 Annotation Mode

In Annotation Mode the user has to locate the closest antecedent markable of an anaphor markable highlighted in orange i.e. an earlier mention of the object. The user can move the cursor over the text and markables are revealed in a bordered box. To select it the user clicks on the bordered box and the markable becomes highlighted in blue. 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 user can 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.

3.3 Validation Mode

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

3.4 Training and Motivating Users

Users 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

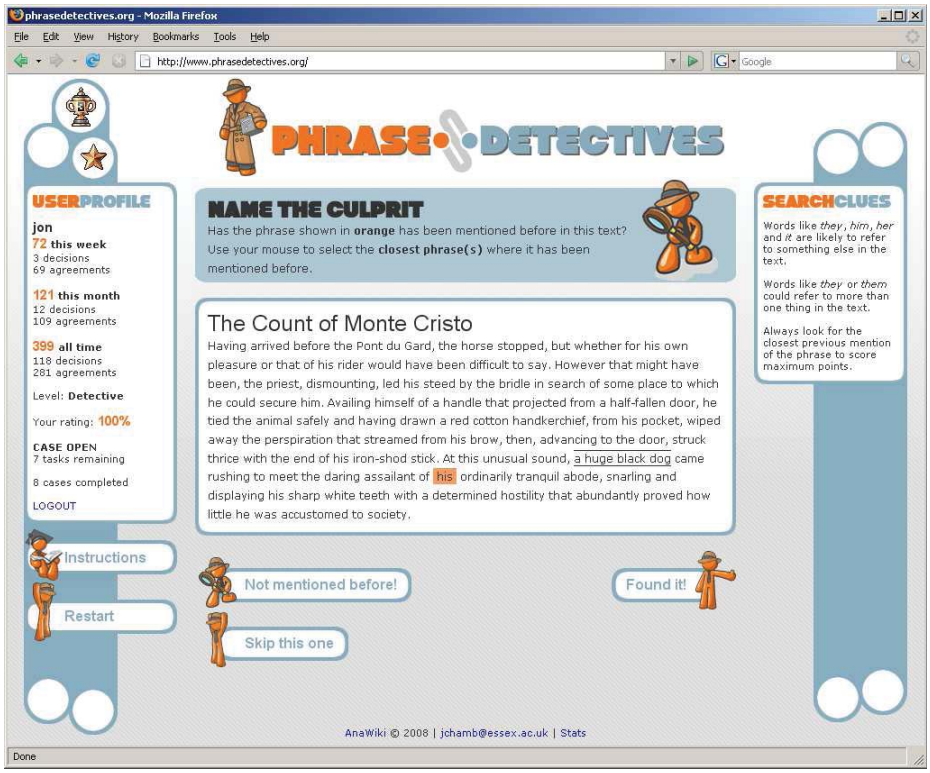

Figure 1: A screenshot of the Annotation Mode.

guidance when they select an incorrect answer and points when they select the correct answer. When the user gives enough correct answers they graduate to annotating texts that will be included in the corpus.

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

The game is designed to motivate users to annotate the text correctly by using comparative scoring (awarding points for agreeing with the Gold Standard), and collaborative scoring (awarding points to the previous user if they are agreed with by the current user). Using leader boards and assigning levels for points has been proven to be an effective motivator, with users often using these as targets [von Ahn, 2006].

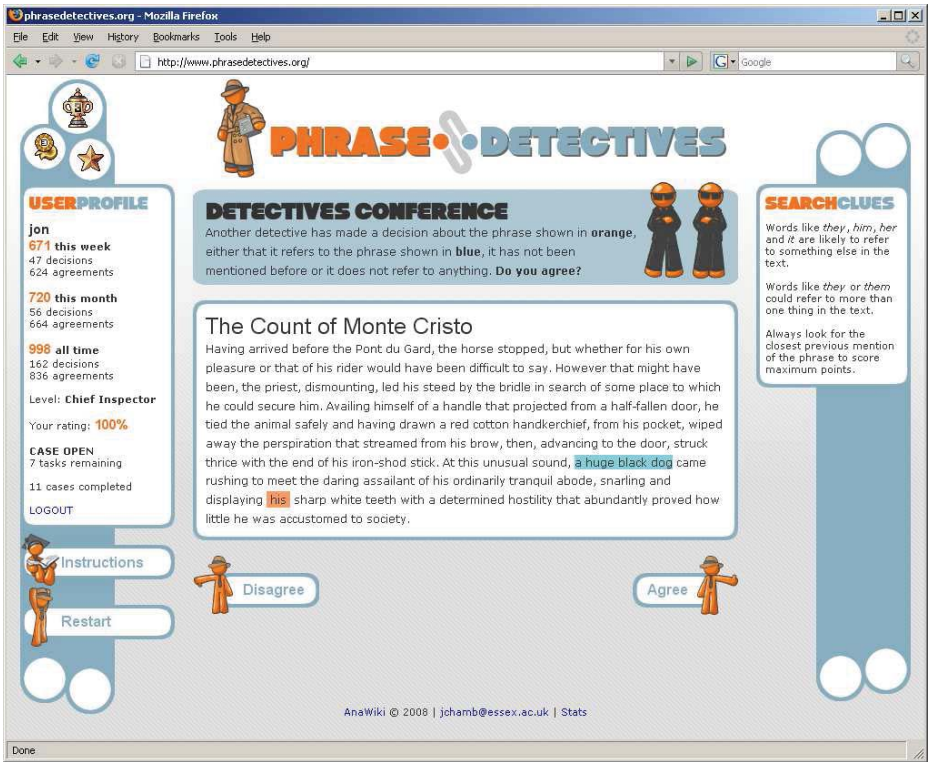

Figure 2: A screenshot of the Validation Mode.

3.5 Preventing Cheating and Filtering Erroneous Annotations

Several methods will be used to identify users who are cheating or who are providing poor annotations. These include checking the IP address, randomly checking annotations against known answers and keeping a blacklist of players to discard all their data [von Ahn, 2006]. Additionally we will time annotations, as this could indicate that the user either did not spend long enough reading the text or it is an automated submission. We anticipate annotation times will be different for each mode, with validation mode being approximately twice as fast as annotation mode [Chklovski and Gil, 2005].

4 Preliminary Study of the Game Interface

A prototype of the game interface was informally evaluated by 16 randomly selected volunteers from the University of Essex which included staff and students.

Feedback was collected in interviews after each session with the aim of getting an insight into the game tasks and the user interface.

We discovered that a training task was necessary, in addition to the instructions, to help the users understand the tasks. Most (80%) of volunteers felt that 2 example tasks would have been sufficient for training.

The reading styles of each volunteer varied considerably, with some reading the whole text, some reading backwards from the markable and others using scanning techniques to look for specific grammatical elements. They were interested in a broad range of topics, including news, travel, factual and literature.

Of the volunteers who used Facebook (67%), all said they would be motivated to play the game if it was integrated with their profile. It is our intention to use social networking sites (including Facebook, Bebo, and MySpace) to attract volunteers to the game and motivate participation by providing widgets (code segments that display the user's score and links to the game) to add to their profile pages.

A beta version of the game was released online in May 2008 to evaluate the game interface, review the systems in place, to train users and determine the quality of the annotations compared to the Gold Standard.

5 Corpus Selection

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. 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. We also expect to include sample text from emails (e.g. from the Enron corpus), text from the American National Corpus and transcripts of spoken text.

The chosen texts will be stripped of all presentation formatting, HTML and links to create the raw text. This will be automatically parsed for POS tags and to extract markables consisting of noun phrases. The resulting XML file can then be inserted into the game database to be annotated.

6 Future Work

Our aim is to have a fully functioning game annotating a corpus of one million words by September 2008. We will be considering extending the interface to include different annotation tasks, for example marking coreference chains or Semantic Web mark-up and will present the game interface to gain feedback from the linguistic and Semantic Web community.

Acknowledgements

ANAWIKI is funded by EPSRC grant number EP/F00575X/1. Thanks to Ron Artstein as well as the Sekimo people at the University of Bielefeld: Daniela Goecke, Maik Stührenberg, Nils Diewald and Dieter Metzing. We also want to thank all volunteers who have already contributed to the project.

References

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

[Chklovski and Gil, 2005] Chklovski, T. and Gil, Y. (2005). Improving the design of intelligent acquisition interfaces for collecting world knowledge from web contributors. In *Proceedings of K-CAP '05*, pages 35–42.

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

[Krötzsch et al., 2007] Krötzsch, M., Vrandečič, D., Völkel, M., Haller, H., and Studer, R. (2007). Semantic Wikipedia. *Journal of Web Semantics*, 5:251–261.

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

[Palmer et al., 2005] Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

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

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

[Poesio et al., 2008] Poesio, M., Kruschwitz, U., and Chamberlain, J. (2008). ANAWIKI: Creating anaphorically annotated resources through Web cooperation. In *Proceedings of LREC'08*, Marrakech.

[Singh, 2002] Singh, P. (2002). 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.

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

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

[von Ahn et al., 2006] von Ahn, L., Liu, R., and Blum, M. (2006). Peekaboom: a game for locating objects in images. In *Proceedings of CHI '06*, pages 55–64.

[Zaenen, 2006] Zaenen, A. (2006). Mark-up Barking Up the Wrong Tree. *Computational Linguistics*, 32(4):577–580.