Addressing the Resource Bottleneck to Create Large-Scale Annotated Texts

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

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

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

Abstract

Large-scale linguistically annotated resources have become available in recent years. This is partly due to sophisticated automatic and semi-automatic approaches that work well on specific tasks such as part-of-speech tagging. For more complex linguistic phenomena like anaphora resolution there are no tools that result in high-quality annotations without massive user intervention. Annotated corpora of the size needed for modern computational linguistics research cannot however be created by small groups of hand annotators. The ANAWIKI project strikes a balance between collecting high-quality annotations from experts and applying a game-like approach to collecting linguistic annotation from the general Web population. More generally, ANAWIKI is a project that explores to what extend expert annotations can be substituted by a critical mass of non-expert judgements.

1 Introduction

Syntactically annotated language resources have long been around, but 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. 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). The ANAWIKI project[1] presents an effort to create high-quality, large-scale anaphorically annotated resources (Poesio et al., 2008) by taking advantage of the collaboration of the Web community, both through cooperative annotation efforts using traditional annotation tools and through the use of game-like interfaces. This makes ANAWIKI a very ambitious project. It is not clear to what extend expert annotations can in fact be substituted by those judgements submitted by the general public as part of a game. If successful, ANAWIKI will actually be more than just an anaphora annotation tool. We see it as a framework aimed at creating large-scale annotated corpora in general.

2 Creating Resources through Web Collaboration

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). Medium and large-scale semantic annotation projects (coreference, wordsense) are a fairly recent innovation in Computational Linguistics (CL). 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

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

enough on general text.

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 (Singh, 2002) to which 14,500 participants contributed nearly 700,000 sentences. A more recent, and perhaps more intriguing, development is the use of interactive game-style interfaces to collect knowledge such as Phetch, Verbosity and Peekaboom (von Ahn et al., 2006). Perhaps the best known example of this approach is the ESP game, a project to label images with tags through a competitive game (von Ahn, 2006); 13,500 users played the game, creating 1.3M labels in 3 months. 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.

ANAWIKI builds on the proposals for marking anaphoric information allowing for ambiguity developed in ARRAU (Poesio and Artstein, 2005) and previous projects. 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.

3 Annotation Tools

Attempts to create hand annotated corpora face the dilemma of either going for the traditional CL approach of high-quality annotation (of limited size) by experts or to involve a large population of non-experts which could result in large-scale corpora of inferior quality. The ANAWIKI project bridges this gap by combining both approaches to annotate the data: an expert annotation tool and a game interface. Both tools are essential parts of ANAWIKI. We briefly describe both, with a particular focus on the game interface.

3.1 Expert Annotation Tool

An expert annotation tool is used to obtain Gold Standard annotations from computational linguists. In the case of anaphora annotation we use the Serengeti tool developed at the University of Bielefeld (Stührenberg et al., 2007). The anaphoric annotation of markables within this environment will be very detailed and will serve as a training corpus as well as quality check for the second tool (see below). Figure 1 is a screenshot of this interface.

3.2 Game Interface

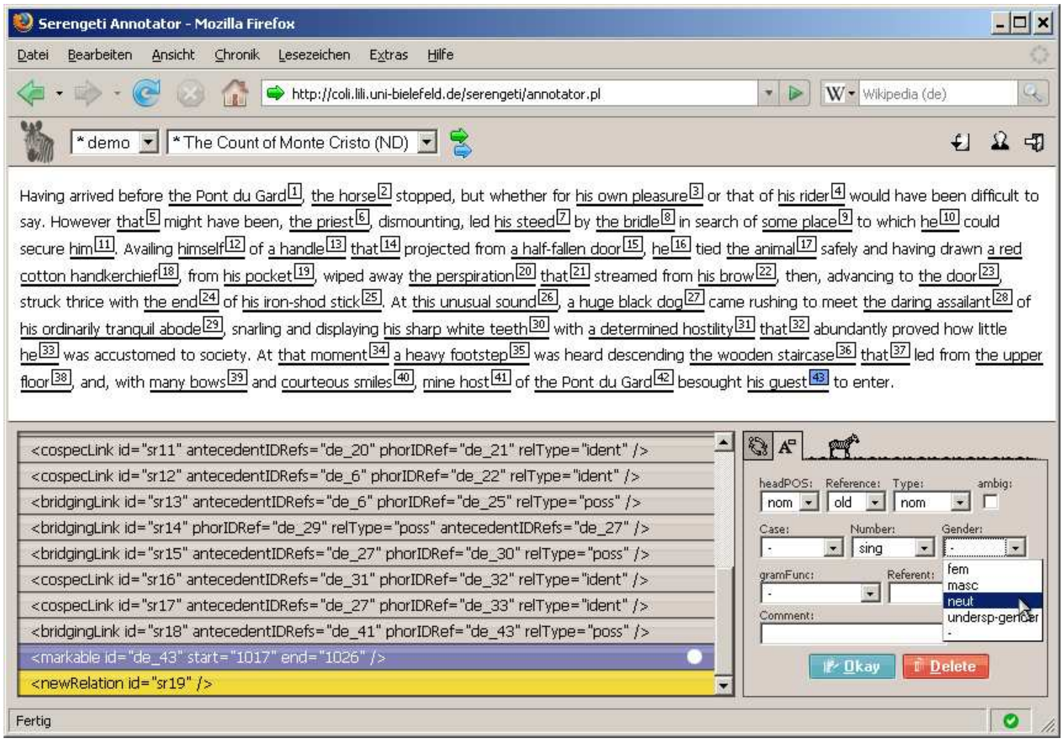
A game interface is used to collect annotations from the general Web population. The game interface integrates with the database of the expert annotation tool but aims to collect large-scale (rather than detailed) anaphoric relations. Users are simply asked to assign an anaphoric link but are not asked to specify what type (or what features) are present.

Phrase Detectives[2] 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 the markable that is the antecedent of the anaphor (Annotation Mode - see Figure 2); or by validating a decision previously submitted by another user (Validation Mode). One motivation for Validation Mode is that we anticipate it to be twice as fast as Annotation Mode (Chklovski and Gil, 2005).

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 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. This is the foundation of the user rating system used 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 retroactive scoring (awarding points to the previous user if they are agreed

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

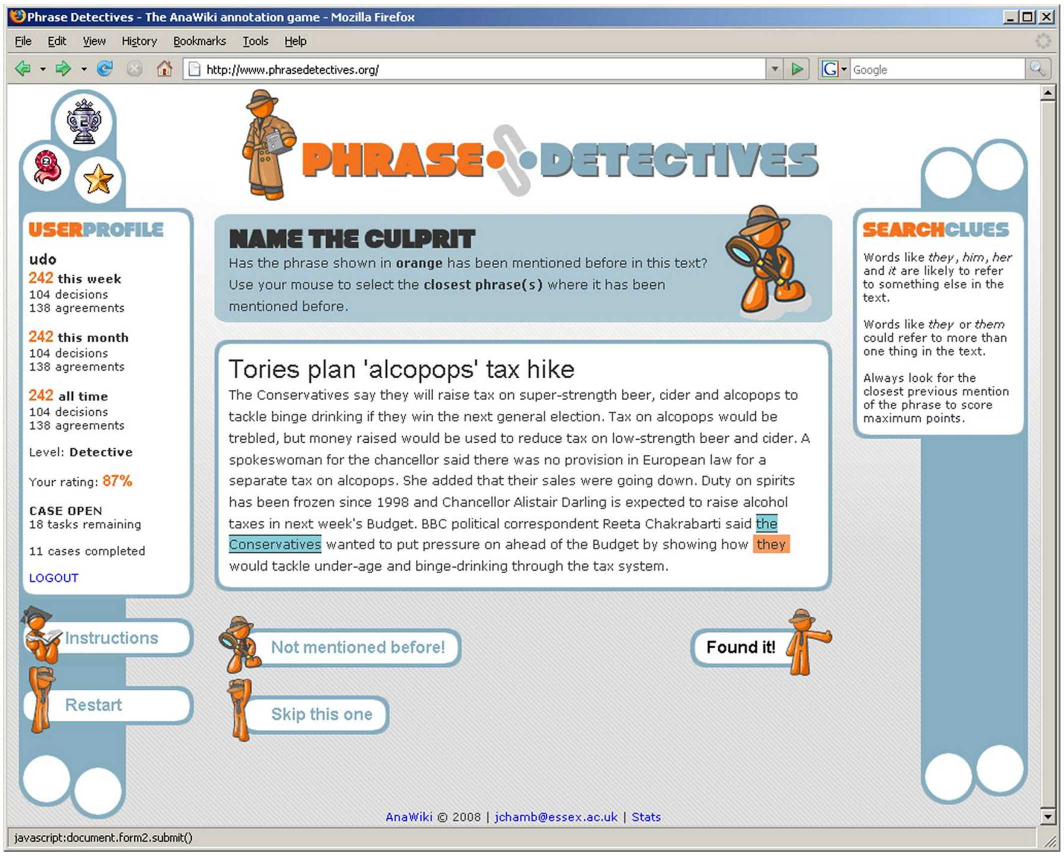

Serengeti Annotator - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
http://coli.lili.uni-bielefeld.de/serengeti/annotator.pl
Wikipedia (de)
* demo
* The Count of Monte Cristo (ND)
Okay
Delete
Fertig

Having arrived before the Pont du Gard[1], the horse[2] stopped, but whether for his own pleasure[3] or that of his rider[4] would have been difficult to say. However that[5] might have been, the priest[6], dismounting, led his steed[7] by the bridle[8] in search of some place[9] to which he[10] could secure him[11]. Availing himself[12] of a handle[13] that[14] projected from a half-fallen door[15], he[16] tied the animal[17] safely and having drawn a red cotton handkerchief[18], from his pocket[19], wiped away the perspiration[20] that[21] streamed from his brow[22], then, advancing to the door[23], struck thrice with the end[24] of his iron-shod stick[25]. At this unusual sound[26], a huge black dog[27] came rushing to meet the daring assailant[28] of his ordinarily tranquil abode[29], snarling and displaying his sharp white teeth[30] with a determined hostility[31] that[32] abundantly proved how little he[33] was accustomed to society. At that moment[34] a heavy footstep[35] was heard descending the wooden staircase[36] that[37] led from the upper floor[38], and, with many bows[39] and courteous smiles[40], mine host[41] of the Pont du Gard[42] besought his guest[43] to enter.

```
<cospecLink id="sr11" antecedentIDRefs="de_20" phorIDRef="de_21" relType="ident" />
<cospecLink id="sr12" antecedentIDRefs="de_6" phorIDRef="de_22" relType="ident" />
<bridgingLink id="sr13" antecedentIDRefs="de_6" phorIDRef="de_25" relType="poss" />
<bridgingLink id="sr14" phorIDRef="de_29" relType="poss" antecedentIDRefs="de_27" />
<bridgingLink id="sr15" antecedentIDRefs="de_27" phorIDRef="de_30" relType="poss" />
<cospecLink id="sr16" antecedentIDRefs="de_31" phorIDRef="de_32" relType="ident" />
<cospecLink id="sr17" antecedentIDRefs="de_27" phorIDRef="de_33" relType="ident" />
<bridgingLink id="sr18" antecedentIDRefs="de_41" phorIDRef="de_43" relType="poss" />
<markable id="de_43" start="1017" end="1026" />
<newRelation id="sr19" />
```

headPOS: nom
Reference: old
Type: nom
ambig: □
Case: -
Number: sing
Gender: -
fem
masc
neut
undersp-gender
-
gramFunc: -
Referent:
Comment:

Figure 1: A screenshot of the Serengeti expert annotation tool.


Phrase Detectives - The AnaWiki annotation game - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.phrasedetectives.org/
Google
PHRASE DETECTIVES
Instructions
Restart
Not mentioned before!
Skip this one
Found it!
AnaWiki © 2008 | jchamb@essex.ac.uk | Stats
javascript:document.form2.submit()

USERPROFILE

udo
242 this week
104 decisions
138 agreements

242 this month
104 decisions
138 agreements

242 all time
104 decisions
138 agreements

Level: Detective

Your rating: 87%

CASE OPEN
18 tasks remaining

11 cases completed

LOGOUT

NAME THE CULPRIT

Has the phrase shown in **orange** has been mentioned before in this text? Use your mouse to select the **closest phrase(s)** where it has been mentioned before.

Tories plan 'alcopops' tax hike

The Conservatives say they will raise tax on super-strength beer, cider and alcopops to tackle binge drinking if they win the next general election. Tax on alcopops would be trebled, but money raised would be used to reduce tax on low-strength beer and cider. A spokeswoman for the chancellor said there was no provision in European law for a separate tax on alcopops. She added that their sales were going down. Duty on spirits has been frozen since 1998 and Chancellor Alistair Darling is expected to raise alcohol taxes in next week's Budget. BBC political correspondent Reeta Chakrabarti said the Conservatives wanted to put pressure on ahead of the Budget by showing how they would tackle under-age and binge-drinking through the tax system.

SEARCHCLUES

Words like *they*, *him*, *her* and *it* are likely to refer to something else in the text.

Words like *they* or *them* could refer to more than one thing in the text.

Always look for the closest previous mention of the phrase to score maximum points.

Figure 2: A screenshot of the Game Interface (Annotation Mode).

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).

The game interface is described in more detail elsewhere (Chamberlain et al., 2008).

4 Challenges

We are aiming at a balanced corpus, similar to the BNC, that includes texts from Project Gutenberg, the Open American National Corpus, the Enron corpus and other freely available sources. The chosen texts are stripped of all presentation formatting, HTML and links to create the raw text. This is automatically parsed to extract markables consisting of noun phrases. The resulting XML format is stored in a relational database that can be used in both the expert annotation tool and the game.

There are a number of challenges remaining in the project. First of all, the fully automated processing of a substantial (i.e. multi-million) word corpus comprising more than just news articles turned out to be non-trivial both in terms of robustness of the processing tools as well as in terms of linguistic quality.

A second challenge is to recruit enough volunteers to annotate a 100 million word corpus within the timescale of the project. 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.

Finally, the project's aim is to generate a sufficiently large collection of annotations from which semantically annotated corpora can be constructed. The usefulness of the created resources can only be proven, for example, by training anaphora resolution algorithms on the resulting annotations. This will be future work.

5 Next Steps

We are currently in the process of building up a critical mass of source texts. Our aim is to have a corpus size of 1M words by September 2008. By this time we also intend having a multilingual user interface (initially English, Italian and German) with the capacity to annotate texts in different languages although this is not the main focus. In the future we will be considering extending the interface to include different annotation tasks, for example marking coreference chains or Semantic Web mark-up. We would like to present the game interface to gain feedback from the linguistic community.

Acknowledgements

ANAWIKI is funded by EPSRC (EP/F00575X/1). Thanks to Daniela Goecke, Maik Stührenberg, Nils Diewald and Dieter Metzing. We also want to thank all volunteers who have already contributed to the project and the reviewers for valuable feedback.

References

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

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

Chklovski, T. and Y. Gil. 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, E., M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. 2006. OntoNotes: The 90% Solution. In *Proceedings of HLT-NAACL06*.

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

Poesio, M. and R. Artstein. 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, M., U. Kruschwitz, and J. Chamberlain. 2008. ANAWIKI: Creating anaphorically annotated resources through Web cooperation. In *Proceedings of LREC'08*, Marrakech.

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

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

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