The Phrase Detective Multilingual Corpus, Release 0.1

Massimo Poesio,[†] Jon Chamberlain,[†] Udo Kruschwitz,[†] Livio Robaldo, Luca Ducceschi

[†]University of Essex, University of Torino, University of Utrecht

Abstract

The Phrase Detectives Game-With-A-Purpose for anaphoric annotation has been live since December 2008, collecting over 2.5 million judgments on the anaphoric expressions in texts in two languages (English and Italian) from around 9,000 players. In this paper we summarize our recent work on creating a corpus using these annotations.

1. Introduction

Phrase Detectives, an interactive online **game with a purpose** (von Ahn, 2006) for creating anaphorically annotated resources making use of a highly distributed population of contributers with different levels of expertise, is an illustration of a new approach for creating large-scale resources: exploiting collective intelligence. In this paper we briefly discuss the language resources side of the enterprise–i.e., how the corpus has been prepared for annotation, the coding scheme, the data being annotated, and the agreement on the annotation.

2. The Game

Phrase Detectives is a single-player game-with-a-purpose developed to collect data about anaphora and centered around the detective metaphor. The game architecture is articulated around a number of **tasks** and uses scoring, progression and a variety of other mechanisms to make the activity enjoyable. A mixture of incentives, from the personal (scoring, levels) to the social (competing for some players, participating in a worthwhile enterprise for others) to the financial (small prizes) are employed.

2.1. Game Design

In *Phrase Detectives* the player is a **detective** that goes about resolving **cases**–expressing judgments about the interpretation of markables–in the so-called **Name-the-Culprit** activity, and providing opinions about other detectives's judgment in the **Detectives Conference** activity. Both of these activities lead to point accumulation, which is the main objective of the players; in fact, as we will see below, validation (Detectives Conference) is the main scoring activity for players once they pass the training threshold.

Name-the-Culprit Name-the-Culprit is the primary activity dedicated to the labelling of data by players. The players are shown a window of text in which a markable is highlighted in orange, as shown in Figure 1 (on the left).[1] They have to decide, first of all, whether the markable is referring, a property, or non-referring. In case they decide the markable is referring, they then have to decide whether it introduces a new entity (i.e., whether it is discourse new), or whether it refers to an already mentioned entity–and in this case they have to locate the closest mention. Moving

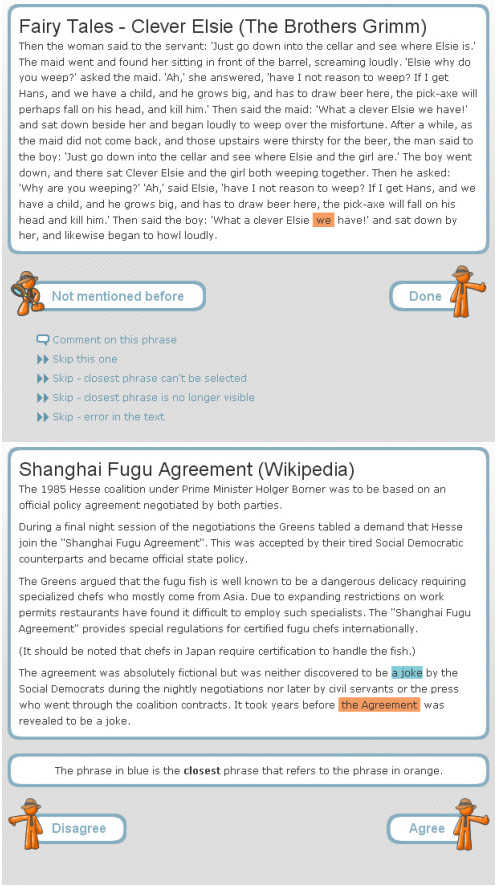

Figure 1: Screenshots of Annotation Mode (top) and Validation Mode (bottom)

the cursor over the text reveals the markables within a bordered box; to select a markable the player clicks on the bordered box and the markable becomes highlighted in blue.

Detectives Conference Every markable for which multiple interpretations have been proposed (the great majority, as discussed in Section 4.) must go through the validation process, **Validation Mode**–aka the **Detectives Conference** activity, displayed on the right side of Figure 1. In Detectives Conference players have to say whether they agree or disagree with an interpretation.

2.2. Other Points

The game-with-a-purpose approach to resource annotation was adopted not just to annotate large amounts of text, but also to collect a large number of judgments about each lin-

[1]These markables are automatically extracted from the text using the pipeline(s) discussed below.

guistic expression, which led to the deployment of a variety of mechanisms for quality control which try to reduce the amount of unusable data beyond those created by malicious users, from the level mechanism itself to validation to a number of tools for analyzing the behavior of players. More recently, a Facebook version was developed.

3. The Corpus

The ultimate goal of Phrase Detectives is to obtain very large anaphorically annotated corpora for the languages covered (currently, English and Italian).

3.1. Coding Scheme

The *Phrase Detectives* corpus is annotated according to the linguistically-oriented approach to anaphoric annotation that is currently prevalent, having been adopted in OntoNotes (Pradhan et al., 2007), our own ARRAU corpus (Poesio and Artstein, 2008) and in all the corpora used in the 2010 SEMEVAL anaphora evaluation (Recasens et al., 2010). In this type of annotation, all NPs are considered markables, and anaphoric relations between all types of entities are annotated, unlike the practice in the MUC and ACE corpora.[2] (E.g., in the *Phrase Detectives* corpora, coordinated NPs like *John and Mary* are also markables.)

Players can assign four types of interpretation (labels) to markables:

- `DN` (discourse-new): the markable refers to a newly introduced entity.
- `DO` (discourse-old): the markable refers to an already mentioned entity; the player has to specify the latest mention.
- `NR` (non-referring): the markable is non-referring (e.g. pleonastic *it*).
- `PR` (property attribute): the markable represents a property of a previously mentioned entity (e.g., *a teacher* in "He is a teacher").

3.2. Input / Output

The data handled by *Phrase Detectives* are stored in a relational database whose design for the part concerned with storing texts and their annotations is based on that of the University of Bielefeld's Serengeti system (Poesio et al., 2011). New texts are entered in the system through the Serengeti interface, that requires input in SGF format (Stührenberg and Goecke, 2008). The text must have been preprocessed to identify tokens, sentences, and noun phrases. The data are outputted in an extended version of the MAS-XML format (Kabadjov, 2007), designed to represent anaphoric information and to encode multiple interpretations. The extended version of MAS-XML, called PD-MAS-XML, can be used to export each interpretation assigned to each markable in the text.

[2] `http://projects.ldc.upenn.edu/ace/data/`

3.3. MAS-XML

The PD-MAS-XML format used to export *Phrase Detectives* data is a modified version of the Minimum Anaphoric Syntax (MAS-XML) format proposed in (Kabadjov, 2007). MAS-XML is a form of inline XML in which the basic information required to carry out resolution is marked, including

- sentences;
- words with their part-of-speech tags (for English, the Penn Treebank tagset is used);
- NPs (called Nominal Entities, `ne`), with their ID and the basic agreement features: gender (attribute `gen` for gold-standard info, `AAgen` for automatically extracted information), number (again two attributes are used, `num` and `AAnum`), and person (using the attributes `per` and `AAper`)
- NP modifiers and heads, using the elements `mod` and `nphead`

Anaphoric information is marked using separate `ante` elements, a structured representation inspired by the Text Encoding Initiative `link` elements and that makes it possible to specify multiple anaphoric relations for each markable (identity and association) and to mark ambiguity using multiple `anchor` elements (Poesio, 2004).

The MAS-XML file for each document that is exported contains the original text and markup (sentences, NPs and their features and constituents) automatically computed by the import pipeline, as well as the annotations produced by the players. To export the annotation information, the anchor mechanism from MAS-XML was replaced by a much more extensive format specifying for every player that expressed a judgment about a given markable the interpretation (`DN` for Discourse-New, `DO` for Discourse-Old, `NR` for Non-Referring, or `PR` for Property), any antecedents selected for `DO` and `PR` interpretations, the user ID, the user rating, the time it took to make the annotation, whether the decision is an agreement and in what mode the decision occurred (annotation or validation). Additionally players' comments are exported with the relevant markable and include the user ID, the type of comment and the text that was submitted; and so are skips. For instance the (real-life) interpretation of markable `ne14817`, which all players interpreted as `DN`, is as follows.

```
<PDante id="ne14817">
   <interpretation>
      <anchor type="DN" user_id="281" user_rating="75"
           annotation_time="2" agree="y" mode="a"/>
      <anchor type="DN" user_id="728" user_rating="58"
           annotation_time="2" agree="y" mode="a"/>
      <anchor type="DN" user_id="779" user_rating="77"
           annotation_time="5" agree="y" mode="a"/>
      <anchor type="DN" user_id="281" user_rating="75"
           annotation_time="1" agree="y" mode="a"/>
      <anchor type="DN" user_id="18" user_rating="77"
           annotation_time="5" agree="y" mode="a"/>
      <anchor type="DN" user_id="1293" user_rating="64"
           annotation_time="15" agree="y" mode="a"/>
      <anchor type="DN" user_id="1364" user_rating="59"
           annotation_time="4" agree="y" mode="a"/>
      <anchor type="DN" user_id="163" user_rating="80"
           annotation_time="2" agree="y" mode="a"/>
      <anchor type="DN" user_id="1659" user_rating="92"
           annotation_time="9" agree="y" mode="a"/>
   </interpretation>
   <skip total="0"/>
</PDante>
```

Documents can be exported from *Phrase Detectives* in MAS-XML format either when they are complete (i.e when all the markables have been annotated sufficiently according to the game configuration) or when they are partially complete. For the purposes of testing only complete documents have been exported.

3.4. Preprocessing

Adding texts in a new language to *Phrase Detectives* requires developing a **pipeline** to convert documents into SGF format importable in the database. Two such pipelines have been developed so far.

The English Pipeline The English *Phrase Detectives* pipeline converting raw text to SGF was developed by combining existing tools (OpenNLP tokenizer and sentence splitter, Berkeley Parser) with *ad-hoc* modules for correcting the output of such tools in the case of frequent errors.

The Italian Pipeline In order to use *Phrase Detectives* to annotate Italian data, a new pipeline (Robaldo et al., 2011) was developed using the TULE parser (Lesmo and Lombardo, 2002). The parser processed the raw text directly with Italian texts so no pre-processing is needed.

3.5. The English and Italian Corpora

As our ultimate goal is to produce a freely distributable corpus, the texts of the English and Italian corpus are from collections not subject to copyright restrictions.

English The English texts come from three main domains:

- Wikipedia articles selected from the 'Featured Articles' page[3] and the page of 'Unusual Articles'[4];
- narrative text from Project Gutenberg[5] including in particular a number of tales (e.g., Aesop's Fables, Grimm's Fairy Tales, Beatrix Potter's tales), and more advanced narratives such as several Sherlock Holmes short stories by A. Conan-Doyle, *Alice in Wonderland*, and several short stories by Charles Dickens.
- dialogue texts from Textfile.[6]

The ultimate objective is to annotate over 100 million words, and several millions words of text have already been converted, but in part because the accuracy of the present pipeline is not considered high enough, at present only around a million words have been actually uploaded in the English version of *Phrase Detectives*–to be precise, 1,206,597 words from 839 documents.

Italian The same criteria concerning distribution were used for the texts in the Italian version of the game; an additional criterion has been the kind of linguistic phenomena that they are likely to include. The sources are the Italian version of Wikipedia and two novels by Wu Ming (CC licensed).

The texts from Wikipedia belong to two specific sub-genres (plots and biographies) which are likely to contain a dense net of antecedents. The first kind displays a significant number of pronominal anaphors, while the second might display examples of lexical noun phrase anaphora (e.g., "the Queen" and "her Majesty.") In addition to the mentioned sub-genres other uncategorized texts have been chosen in order to provide a comparison with the English version of the game ("Chess Boxing" and "Diet Coke and Mentos Explosion" are in both corpora).

The novels have been selected to test if the narrative style has an influence on the performance of the parser and of the players. This variety is more likely to display all the pronouns of the language, particularly 1st and 2nd person in reported speech, which are less likely to appear in Wikipedia articles.

The Italian corpus for *Phrase Detectives* currently contains 30 texts, for a total of 11,373 words.

4. Results So Far

4.1. A Quantitative Assessment

Since the first release of the game in December 2008 to January 2012 just over 10,000 players have registered (10,250 as this paper is completed), 2,000 of which went beyond the initial training phase. 665 of these players are using the Facebook version.

445 documents have been fully annotated, for a total completed corpus of 181,000 words, 15% of the total size of the collection currently uploaded for annotation in the game (1.2M words). This is comparable in size to the ACE2 corpus of anaphoric information (BNews + Npaper + Nwire),[7], which was the standard for evaluation of anaphora resolution systems until 2007/08 and still widely used. The size of the completed corpus does not properly reflect, however, the amount of data we have collected, as the case allocation strategy adopted in the game privileges variety over completion rate; as a result, almost all the 841 documents in the corpus have already been partially annotated. This is reflected, e.g., in the fact that 84280 of the 392,120 markables in the active documents (21%) have already been annotated. This is already almost twice the total number of markables in the entire OntoNotes 3.0 corpus,[8] which contains 1 million tokens, but only 45,000 markables.

4.2. Agreement on Annotations

In order to check the extent to which the annotations produced by the game corresponded to the annotations produced by experts, we randomly selected five completed documents from the Wikipedia corpus containing 154 markables. Each document was manually annotated by two experts (called Expert 1 and Expert 2 in the rest of this discussion) operating separately; we then compared the annotations produced by the experts with the most highly

[3] http://en.wikipedia.org/wiki/Wikipedia:Featured_articles

[4] http://en.wikipedia.org/wiki/Wikipedia:Unusual_articles

[5] http://www.gutenberg.org/

[6] http://www.textfiles.com/

[7] http://projects.ldc.upenn.edu/ace/data/

[8] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2009T24

ranked interpretations produced by the players (henceforth, the **game interpretation**), and with each other.

Overall, agreement between experts on the types is very high although not complete: 94%, for a chance-adjusted κ value (Artstein and Poesio, 2008) of $\kappa = .87$, which is extremely good. This value can be seen as an upper boundary on what we might get out of the game. Agreement between each of the experts and the majority interpretation of the game is also good: we found 84.5% percentage agreement between Expert 1 and the game ($\kappa = .71$) and 83.9% agreement between Expert 2 and the game ($\kappa = .7$). In other words, in about 84% of all cases the interpretation specified by the majority vote of non-experts was identical to the one assigned by an expert. These values are comparable to those obtained when comparing an expert with the 'normally trained' annotators (usually students) that are typically used to create medium-quality resources.

4.3. Ambiguity in the Corpus

We are in the process of analyzing the judgments accumulated so far in preparation for a paper on anaphora through the lens of *Phrase Detectives*, and some interesting results already came up, in particular about the notion of coreference (e.g., in many mysteries, the whole point of the story is that the identity of a character–the culprit, or some shady figure– is only discovered at the end). We will not enter into this discussion here, but one preliminary statistic is worth reporting given the motivating role that studying anaphoric ambiguity has had in the design of the game. In January 2011 there were 63009 completely annotated markables. Of these, 23479 (37.3%) had exactly one interpretation (i.e., the first eight players to be presented with that markable all chose the same interpretation). Of these, 23,138 were DN, 322 DO, and 19 NR. A further 13,772 markables (21%) had only 1 interpretation with a score greater than 0. Again, the majority of these (9,194) were DN; 4,391 were DO, and NR 175.

5. Discussion

Phrase Detectives was one of the very first GWAP applied to resource creation for HLT and in quantitative terms has been the most successful, collecting over 2.5 million judgments from over 10,000 players. Annotation is still going strong and we expect it to continue for the immediate future; our hope is to complete at least the annotation of the initial 1.2M corpus of documents. In order to annotate more data, a higher-quality preprocessing pipeline for English will be required.

Among the lessons we learned, the first and most obvious is that GWAP can be used for HLT resource creation. However researchers will need to consider with great care whether in fact this approach is appropriate for their task and their data. If only a small amount of data is required (100,000 words or less), and / or the data are not very interesting, it may be best to use crowdsourcing instead. If the GWAP approach is chosen, a constant effort of promotion will be required to make the game stand out among the thousands of other games (serious or not)–but offering small prizes proved very effective.

Concerning the architecture of the game, the main lesson we learned is that validation is essential and very effective for quality control. Keeping around all interpretations also proved the right choice. Last but not least, embedding the game in Facebook has proven very effective not so much as a new way of reaching players but to know better who your players are.

Next steps include developing methods for cleaning up the data and for using the data to train anaphoric models.

6. Acknowledgements

The initial funding for *Phrase Detectives* (2007/09) came from EPSRC project AnaWiki, EP/F00575X/1.

7. References

R. Artstein and M. Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

M. A. Kabadjov. 2007. *Task-oriented evaluation of anaphora resolution*. Ph.D. thesis, University of Essex.

L. Lesmo and V. Lombardo. 2002. Transformed subcategorization frames in chunk parsing. In *Proc. of LREC*, pages 512–519, Las Palmas.

M. Poesio and R. Artstein. 2008. Anaphoric annotation in the arrau corpus. In *Proc. of LREC*, Marrakesh, May.

M. Poesio, N. Diewald, M. Stührenberg, J. Chamberlain, D. Jettka, D. Goecke, and U. Kruschwitz. 2011. Markup infrastructure for the anaphoric bank: Supporting web collaboration. In A.Mehler, K.-U. Kühnberger, H. Lobin, H. Lüngen, A. Storrer, and A. Witt, editors, *Modeling, Learning, and Processing of Text Technological Data Structures*, Springer, pages 175–195.

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

S. S. Pradhan, L. Ramshaw, R. Weischedel, J. MacBride, and L. Micciulla. 2007. Unrestricted coreference: Indentifying entities and events in ontonotes. In *Proc. ICSC*, Irvine, CA.

M. Recasens, L. Màrquez, E. Sapena, M. A. Martí, M. Taulé, V. Hoste, M. Poesio, and Y. Versley. 2010. Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proc. SEMEVAL*, Uppsala.

L. Robaldo, M. Poesio, L. Ducceschi, J. Chamberlain, and U. Kruschwitz. 2011. Italian anaphoric annotation with the Phrase Detectives game-with-a-purpose. In *Proc. of AIIA*, Lecture Notes in Artificial Intelligence, pages 407–412, Berlin. Springer.

M. Stührenberg and D. Goecke. 2008. SGF – An integrated model for multiple annotations and its application in a linguistic domain. In *Balisage: The Markup Conference*, Montreal, Kanada.

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