Phrase Detectives: Utilizing Collective Intelligence for Internet-Scale Language Resource Creation (Extended Abstract)*

Massimo Poesio†, Jon Chamberlain†, Udo Kruschwitz†, Livio Robaldo‡ and Luca Ducceschi§

Abstract

We are witnessing a paradigm shift in human language technology that may well have an impact on the field comparable to the statistical revolution: acquiring large-scale resources by exploiting collective intelligence. An illustration of this approach is *Phrase Detectives*, an interactive online game-with-a-purpose for creating anaphorically annotated resources that makes use of a highly distributed population of contributors with different levels of expertise.

The paper gives an overview of all aspects of *Phrase Detectives*, from the design of the game and the methods used, to the results obtained so far. It furthermore summarises the lessons that have been learnt in developing the game to help other researchers assess and implement the approach.

1 Introduction

The statistical revolution in Human Language Technology (HLT) has resulted in the first HLT components and applications truly usable on a large scale. It also created, however, a need for large amounts of annotated linguistic data for training and evaluating such systems.

Wikipedia is now routinely used as a word sense repository, possibly even more than WordNet [Csomai and Mihalcea, 2008] or as a source of encyclopedic knowledge [Ponzetto and Strube, 2007]; and **crowdsourcing** through Amazon's Mechanical Turk[1] has quickly become the method of choice for the fast annotation of small-and-not-so-small corpora, and for some types of HLT system evaluation [Snow *et al.*, 2008; Callison-Burch, 2009]. Less used, so far, is an approach to collaborative resource construction: incentivising users to create resources by developing a so-called **game-with-a-purpose** (GWAP) which will produce the required resource as a by-product of the users playing. It is estimated that every year over 9 billion person-hours are spent by people playing games on the Web [von Ahn, 2006]. If even a fraction of this effort could be redirected towards resource creation via the development of Web games we would have enormous quantity of man-hours at our disposal.

The GWAP approach has been used for many different types of crowdsourced data collection including text, image, video and audio annotation, biomedical applications, transcription, search results and social bookmarking [Chamberlain *et al.*, 2013a]. In the paper [Poesio *et al.*, 2013] we discuss *Phrase Detectives*, one of the first GWAP for corpus annotation, and one of the very few such games to result in the collection of a substantial amount of data. The paper is intended to be the definitive reference article on *Phrase Detectives*, collecting in a single publication material previously only found in separate papers, as well as additional material not presented before: a cost comparison between games, traditional annotation and crowdsourcing; a discussion of recent developments such as the Facebook version of the game; and the release of some of the completed data to the research community.

2 Designing Games with a Purpose

As with other methods of crowdsourcing, games should be designed so that the interface is easy to use and the task presented in a way that is simple to understand. Additional considerations for this type of approach include finding a way to **attract** players, and then **incentivise** them to keep playing either by making the task fun or by stimulating their competitive spirit.

Quality control must be considered to prevent malicious users or users who simply do not care or do not understand the underlying rules of the game making the collected data unusable. Controlling cheating may be one of the most important factors in game design. If a player is motivated to work hard and score points, they may become more motivated to find a way to cheat the system. Obtaining reliable results from non-experts is a challenge for crowdsourcing generally and, in this context, strategies for dealing with the issue have been discussed extensively [Kazai *et al.*, 2009; Feng *et al.*, 2009]. Further details of methods to engage and motivate users of human computation systems is presented elsewhere [Chamberlain *et al.*, 2013b].

†[poesio,jchamb,udo]@essex.ac.uk, University of Essex

‡robaldo@di.unito.it, University of Turin

§l.ducceschi@uu.nl, University of Utrecht/Verona

[1] http://www.mturk.com

Figure 1: Screenshot of the *Phrase Detectives* player page.

3 A Game-with-a-Purpose for Language Annotation: Phrase Detectives

Phrase Detectives[2] is a single-player game-with-a-purpose developed to collect data about anaphora coreference[3] [Garnham, 2001; Poesio *et al.*, 2011] and centred around a 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 points, gaining levels) to the social (competing with other players) to the financial (small prizes) were employed. The GWAP approach to resource creation was adopted not just to annotate large amounts of text, but also to collect a large number of judgements about each linguistic expression.

A key decision in the design of *Phrase Detectives* was to follow the approach to data collection adopted in *LEARNER* [Chklovski and Gil, 2005] – namely, to have the Web collaborators perform both the task of providing the judgements (**annotation**) and the task of checking those judgements (**validation**). The inclusion of the latter step plays a crucial role in the strategy for quality control.

In *Phrase Detectives* the player is a detective that goes about resolving cases (expressing judgements about the interpretation of markables) in the so-called **Name-the-Culprit** activity, and providing opinions about other detectives' judgements in the **Detectives Conference** activity. Both of these activities lead to point accumulation, which is the main objective of the players in the game (see Figure 1).

[2] `https://www.phrasedetectives.org`

[3] Anaphoric coreference is a type of linguistic reference where one expression depends on another referential element. An example would be the relation between the entity *'Jon'* and the pronoun *'his'* in the text '*Jon rode his bike to school.*'

Each markable (a segment of text) in a document is presented to several players in Annotation Mode (see Figure 2). If every player chooses the same interpretation then that markable is classified as complete. Every markable for which multiple interpretations have been proposed must go through the validation process. The Annotation-Validation model is explained in more detail elsewhere [Chamberlain, 2014].

Phrase Detectives features the **incentives** usually found in online games for players motivated by a competitive spirit, such as weekly, monthly and all-time leaderboards, cups for monthly top scores and named levels for reaching a certain number of points.

Additionally, monthly **financial incentives** (prizes) for the highest-scoring players in the form of Amazon vouchers sent by email to the winners have been offered regularly since the launch of the game.

From the beginning of the project the choice of documents was considered important in getting players to enjoy the game, to understand the tasks and to keep playing. The documents consist for the most part of narrative texts from the Gutenberg collection and encyclopedic texts from Wikipedia, as well as some existing corpora, television scripts and documents contributed by the players.

The strategies for **quality control** in *Phrase Detectives* address four main issues:

- Training and evaluating players
- Attention slips
- Malicious behaviour
- Multiple judgements and genuine ambiguity

Validation information has proven very effective at identifying interpretations produced by sloppy or malicious players: the value obtained by combining the player annotations with the validations for each interpretation,

$$Ann + Agr - Disagr,$$

(where Ann is the number of players initially choosing the interpretation in Annotation Mode, Agr is the number of players agreeing with that interpretation in Validation Mode, and $Disagr$ is the number of players disagreeing with it in Validation Mode) tends to be zero or negative for all spurious interpretations.

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], the ARRAU corpus [Poesio and Artstein, 2008] and in all the corpora used in the 2010 SEMEVAL anaphora evaluation [Recasens *et al.*, 2010].

The ultimate objective is to annotate over 100 million words and several million 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.

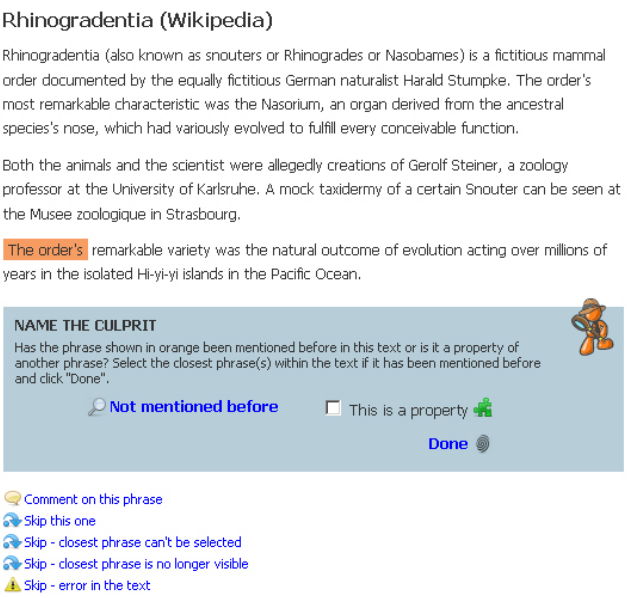

Figure 2: Screenshot of Annotation Mode.

4 Evaluation

The paper reports the results of several forms of evaluation of the results obtained with *Phrase Detectives*: from a quantitative perspective (how many players were recruited, how much labelling they did), as well as from the perspective of the quality of the results, evaluated by **agreement** (how the aggregated results obtained from the game compare to expert judgements) and by using the data to **train anaphoric resolvers**. Additionally we evaluated the **cost-effectiveness** of *Phrase Detectives* in comparison with other types of annotation methods.

4.1 Quantity of data collected

Since the first release of the game in December 2008 to January 2012 just over 8,000 players have registered, of which 3,000 went beyond the initial training phase. These players did more than 5,000 hours of work, i.e., 2.5 person-years. The average throughput of the game (labels per hour) is 450 annotations per hour. Average lifetime play (time in minutes spent on average by players in front of the game summing up all their interactions) is 2,105 secs (35 mins and 5 secs) but this masks a massive difference between players that spend little or no time on the game and players that play continuously.

407 documents were fully annotated, for a total completed corpus of over 162,000 words, 13% of the total size of the collection currently uploaded for annotation in the game (1.2 million words). This is about the size of the ACE2 corpus of anaphoric information, 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 collected, as the task allocation strategy adopted in the game privileges variety over completion rate. As a result, almost all of the documents in the corpus have been partially annotated. This is reflected in the fact that 21% of the 392,120 markables in the active documents have already been fully annotated.

4.2 Quality of the data

One way to tell whether the game is successful at obtaining good quality anaphoric annotations is to check how the aggregated annotations produced by the game compare to those produced by an expert annotator.

Five completed documents (containing 154 markables) were selected from the Wikipedia corpus. Each document was manually annotated by two experts operating separately. The annotations produced by the experts were then compared with the most highly ranked interpretations produced by the players (henceforth, the **game interpretation**), and the experts' annotations with each other.

The experts judged discourse-new (DN) to be the most common interpretation, with 70% of all markables falling in this category. 20% of markables were discourse-old (DO) and form a coreference chain with previous markables. Less than 1% of markables were non-referring (NR) and the remaining markables have been identified as expressing properties (PR).

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 game is also good: 84.5% percentage agreement between Expert 1 and the game ($\kappa = .71$) and 83.9% agreement between Expert 2 and the game ($\kappa = .70$). 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.

An alternative way to evaluate the quality of the annotated data is to use the *Phrase Detectives* data for training anaphora resolution algorithms. We carried out two studies of this type as sanity checks, one in 2009, and one in 2010. In both tests, the BART anaphora resolution toolkit [Broscheit *et al.*, 2010] was used to train a Soon-et-al style model [Soon *et al.*, 2001] by using the top interpretation from the game.

In the first study 21 documents from the Gutenberg corpus were used (a total of 12K words, about half the size of the commonly used MUC6 corpus). 16 documents were used for training and 5 for testing. The performance of the model was measured using the commonly used MUC score [Vilain *et al.*, 1995] which measures precision, recall and F-value at finding anaphoric links. The model achieved F=.58, i.e., on the higher end for systems doing the 'all mentions', 'all modifier' task.

In the second study five times as much data was used from the Wikipedia corpus. 190 documents were used, of which 130 were used for training and 60 for testing (a total of 60K words, about twice the size of the MUC6 corpus). This time the model achieved an F-value of .49, i.e., on the lower end of the performance for this type of dataset.

	Cost (US$)/markable
Traditional, High Quality	3
Traditional, Medium Quality	1.2
Amazon Mechanical Turk	1.2-1.3
Games with a Purpose	.47

Table 1: Comparison of estimated costs (in US$) using four different annotation methods.

4.3 Cost-Effectiveness

One of the main reasons for using GWAP for annotation is the hope that this approach will result in much lower costs in comparison with traditional annotation. In this section we analyse our experience with *Phrase Detectives* from this perspective, also comparing estimated costs of annotation using crowdsourcing (see Table 1). Costs have been converted and expressed as US$ unless otherwise stated.

For **Traditional, High Quality (THQ)** annotation, a formal coding scheme is developed, and often extensive agreement studies are carried out; then every document is doubly annotated according to the coding scheme by two professional annotators under the supervision of an expert, typically a linguist, and annotation is followed by merging of the annotations. It is this type of annotation which requires in the order of $1 million per 1 million tokens, i.e., $1 per token. On average our texts contain around 1 markable every 3 tokens, so we get a cost of $3 per markable.

Traditional, Medium Quality (TMQ) annotation is typically carried out by trained but not professional annotators, generally students, under the supervision of an expert annotator. Our own estimates (at UK / Italy costs) for this type of work are in the order of €330,000/$400,000 per 1 million tokens, including expert annotator costs, i.e., around $0.4 per token, or $1.2 per markable, which is slightly more than 2/5ths of the costs with THQ.

Costs with **crowdsourcing** via Amazon Mechanical Turk depend on the amount paid per HIT and on the extent of duplication and redundancy. In our experience, $0.05 per HIT is the minimum required for non-trivial tasks, and for a task like anaphora, the cost is more like $0.1 per markable. Also, although many researchers only require five judgements per item, in practice we find that 10 is more like the number needed. This results in a cost of $1 per markable, i.e., around $330,000 per million tokens. In addition, an administrator is typically required to set up the task and follow it up. This would give a total cost in the range of $380,000–430,000 per million tokens / $1.2–1.3 per markable, which is about the same cost as with TMQ.

The total cost for running ***Phrase Detectives*** so far has been around £60,000 in salary costs for setting up and running the game and around £6,000 in prizes for three years, i.e., a total of around $100,000 per 162,000 complete tokens. However, over 84,000 markables have been completely annotated, at a cost of $1.2 per markable. With GWAP, most of the expense takes place at the beginning, to set up the game: we spent $65,000 for the first two years at the end of which we had the game, but less than 60,000 words and 10,000 markables fully annotated. In the following 2 years, during which 74,000 markables were completely annotated, the cost has been $35,000 ($0.47 per markable). This figure gives a projected cost of $217,927 for 1 million words ($65,000 for the first 10,000 markables, $152,927 for the other 323,333 markables that one can expect to find on average in 1 million words). This is about half the cost estimated for Amazon Mechanical Turk.

The one problem is that the rate of 34,000 completed markables a year is not fast enough: at that speed it would take 9 years to complete the 307,480 markables remaining in the documents already active in *Phrase Detectives*.

4.4 Ongoing Work and Future Developments

The original Web game is still being played and new data accumulated, but the rate of new data creation and new players registering needs to increase. Therefore, we are continuously launching new initiatives aimed at reaching the numbers we are hoping to achieve.

Phrase Detectives on Facebook The success of games integrated into social networking sites such as Sentiment Quiz on Facebook indicates that visible social interaction within a game environment motivates players to contribute more [Rafelsberger and Scharl, 2009]. This was one of the motivations for developing a Facebook version of *Phrase Detectives*.

Facebook *Phrase Detectives*[4] maintains the overall game architecture whilst incorporating a number of new features developed specifically for the social network platform. The game was developed in PHP SDK (a Facebook API language allowing access to user data, friend lists, and wall posting) and integrates seamlessly within the Facebook site. Data generated from this version of the game is compatible with previous versions and both current implementations of the game run simultaneously on the same corpus.

The game makes full use of socially motivating factors inherent in the Facebook platform, for example any of the player's friends who are playing the game form the player's team and whenever a player's decision agrees with a team member they score additional points.

Post-processing to Improve Quality Ongoing research into data collected from *Phrase Detectives* shows significant improvements in data quality can be achieved with post-processing using different aggregation techniques as well as physical indicators of a player's abilities and attention span to detect outliers and poor performance [Chamberlain and O'Reilly, 2014].

Dataset Availability A sub-corpus of completed documents, including all source text, gold standard annotation and data collected by the game, will be made available to the research community in the near future via the Linguistic Data Consortium (LDC)[5] and the Anaphoric Bank.[6]

[4] https://apps.facebook.com/phrasedetectives

[5] https://www.ldc.upenn.edu

[6] https://www.anaphoricbank.org

References

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

[Broscheit *et al.*, 2010] S. Broscheit, M. Poesio, S.-P. Ponzetto, K. J. Rodriguez, L. Romano, O. Uryupina, Y. Versley, and R. Zanoli. BART: A multilingual anaphora resolution system. In *Proceedings of Semantic Evaluation (SEMEVAL) Workshop*, Uppsala, 2010.

[Callison-Burch, 2009] C. Callison-Burch. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*, pages 286–295. Association for Computational Linguistics, 2009.

[Chamberlain and O'Reilly, 2014] J. Chamberlain and C. O'Reilly. User performance indicators in task-based data collection systems. In *Proceedings of MindTheGap'14*, Berlin, 2014.

[Chamberlain *et al.*, 2013a] J. Chamberlain, K. Fort, U. Kruschwitz, M. Lafourcade, and M. Poesio. Using games to create language resources: Successes and limitations of the approach. In *ACM Transactions on Interactive Intelligent Systems*, volume The People's Web Meets NLP: Collaboratively Constructed Language Resources. Springer, 2013.

[Chamberlain *et al.*, 2013b] J. Chamberlain, U. Kruschwitz, and M. Poesio. Methods for engaging and evaluating users of human computation systems. In *Handbook of Human Computation*. Springer, 2013.

[Chamberlain, 2014] J. Chamberlain. The Annotation-Validation (AV) model: Rewarding contribution using retrospective agreement. In *Proceedings of GamifIR'14*, Amsterdam, 2014.

[Chklovski and Gil, 2005] T. Chklovski and Y. Gil. Improving the design of intelligent acquisition interfaces for collecting world knowledge from web contributors. In *Proceedings of the 3rd international conference on Knowledge capture*, pages 35–42, 2005.

[Csomai and Mihalcea, 2008] A. Csomai and R. Mihalcea. Linking documents to encyclopedic knowledge. *IEEE Intelligent Systems*, 2008. Special issue on Natural Language Processing for the Web.

[Feng *et al.*, 2009] D. Feng, S. Besana, and R. Zajac. Acquiring high quality non-expert knowledge from on-demand workforce. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, People's Web '09, pages 51–56, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[Garnham, 2001] A. Garnham. *Mental models and the interpretation of anaphora*. Psychology Press, 2001.

[Kazai *et al.*, 2009] G. Kazai, N. Milic-Frayling, and J. Costello. Towards methods for the collective gathering and quality control of relevance assessments. In *Proceedings of the 32nd international Special Interest Group on Information Retrieval (SIGIR) conference on Research and development in information retrieval*, SIGIR '09, pages 452–459, New York, NY, USA, 2009. ACM.

[Poesio and Artstein, 2008] M. Poesio and R. Artstein. Anaphoric annotation in the ARRAU corpus. In *Proceedings of the sixth International Conference on Language Resources and Evaluation*, Marrakesh, 2008.

[Poesio *et al.*, 2011] M. Poesio, R. Stuckardt, and Y. Versley. *Anaphora Resolution: Algorithms, Resources and Applications*. Springer, Berlin, 2011.

[Poesio *et al.*, 2013] M. Poesio, J. Chamberlain, U. Kruschwitz, L. Robaldo, and L. Ducceschi. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Interactive Intelligent Systems*, 2013.

[Ponzetto and Strube, 2007] S. Ponzetto and M. Strube. Knowledge derived from Wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence Research*, 30:181–212, 2007.

[Pradhan *et al.*, 2007] S. S. Pradhan, L. Ramshaw, R. Weischedel, J. MacBride, and L. Micciulla. Unrestricted coreference: Identifying entities and events in OntoNotes. In *Proceedings of the International Conference on Semantic Computing*, Irvine, CA, 2007.

[Rafelsberger and Scharl, 2009] W. Rafelsberger and A. Scharl. Games with a purpose for social networking platforms. In *Proceedings of the 20th Association for Computing Machinery (ACM) conference on Hypertext and hypermedia*, pages 193–198. ACM, 2009.

[Recasens *et al.*, 2010] M. Recasens, L. Màrquez, E. Sapena, M. A. Martí, M. Taulé, V. Hoste, M. Poesio, and Y. Versley. SemEval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of Semantic Evaluation (SEMEVAL) Workshop*, Uppsala, 2010.

[Snow *et al.*, 2008] R. Snow, B. O'Connor, D. Jurafsky, and A. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Morristown, NJ, USA, 2008. ACL.

[Soon *et al.*, 2001] W. M. Soon, D. C. Y. Lim, and H. T. Ng. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4), 2001.

[Vilain *et al.*, 1995] M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. A model-theoretic coreference scoring scheme. In *Proceedings of the Sixth Message Understanding Conference*, pages 45–52, 1995.

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