A Crowdsourced Corpus of Multiple Judgments and Disagreement on Anaphoric Interpretation

Massimo Poesio  
Queen Mary University  
m.poesio@qmul.ac.uk

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

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

Silviu Paun  
Queen Mary University  
s.paun@qmul.ac.uk

Alexandra Uma  
Queen Mary University  
a.n.uma@qmul.ac.uk

Juntao Yu  
Queen Mary University  
juntao.cn@gmail.com

Abstract

We present a corpus of anaphoric information (coreference) crowdsourced through a game-with-a-purpose. The corpus, containing annotations for about 108,000 markables, is one of the largest corpora for coreference for English, and one of the largest crowdsourced NLP corpora, but its main feature is the large number of judgments per markable: 20 on average, and over 2.2M in total. This characteristic makes the corpus a unique resource for the study of disagreements on anaphoric interpretation. A second distinctive feature is its rich annotation scheme, covering singletons, expletives, and split-antecedent plurals. Finally, the corpus also comes with labels inferred using a recently proposed probabilistic model of annotation for coreference. The labels are of high quality and make it possible to successfully train a state of the art coreference resolver, including training on singletons and non-referring expressions. The annotation model can also result in more than one label, or no label, being proposed for a markable, thus serving as a baseline method for automatically identifying ambiguous markables. A preliminary analysis of the results is presented.

1 Introduction

A number of datasets for anaphora resolution / coreference now exist (Poesio et al., 2016), including ONTONOTES that has been the de facto standard since the CONLL shared tasks in 2011 and 2012 (Pradhan et al., 2012), and the just introduced and very substantial PreCo corpus (Chen et al., 2018). None of these datasets however take into account the research challenging the idea that a ‘gold standard’ interpretation can be obtained through adjudication, in particular for anaphora (Poesio and Artstein, 2005b; Wong and Lee, 2013; Aroyo and Welty, 2015). Virtually every project devoted to large-scale annotation of discourse or semantic phenomena has reached the conclusion that genuine disagreements are widespread. This has long been known for anaphora (Poesio and Artstein, 2005b; Versley, 2008; Recasens et al., 2011) (see also the analysis of disagreements in ONTONOTES in (Pradhan et al., 2012)) and wordsenses (Passonneau et al., 2012), but more recent work has provided evidence that disagreements are frequent for virtually every aspect of language interpretation, not just in subjective tasks such as sentiment analysis (Kenyon-Dean et al., 2018), but even in the case of tasks such as part-of-speech tagging (Plank et al., 2014). In fact, researchers in the CrowdTruth project view disagreement as positive, arguing that “disagreement is signal, not noise” (Aroyo and Welty, 2015). In this paper we present what to our knowledge is the largest corpus containing alternative anaphoric judgments: 20.6 judgments per markable on average (up to 90 judgments in some cases) for about 108,000 markables. We are not aware of any comparable resource for studying disagreement and ambiguity in anaphora or indeed any other area of NLP. We present some preliminary analysis in the paper.

The corpus presented in this paper is also the largest corpus for anaphora / coreference entirely created through crowdsourcing, and one of the largest corpus of coreference information for English in terms of markables. So far, only fairly small coreference corpora have been created using crowdsourcing (Chamberlain et al.; Guha et al., 2015). The corpus presented here provides annotations for about 108,000 markables, 55% of the number of markables in ONTONOTES. Another novelty is that the corpus was created through a ‘quasi’ Game-With-A-Purpose (GWAP) (von Ahn, 2006; Lafourcade et al., 2015), Phrase Detectives (Poesio et al., 2013), and is, to our knowledge, the largest GWAP-created corpus for NLP. (So far, the success of GWAPs in other areas of science (Clery,
2011; Cooper et al., 2010) has not been replicated in NLP. Finally, the corpus is notable for a richer annotation scheme than the other large coreference corpora. Singletons were marked as well as mentions participating in coreference chains (the omission of singletons being one of the main problems with ONTONOTES). Non-referring expressions were also annotated: both expletives (not annotated either in ONTONOTES or PRECO) and predicative NPs. Finally, all types of plurals were annotated, including also split-antecedent plurals as in John met with Mary, and they went to dinner, which again are not annotated either in ONTONOTES or PRECO.

Turning a crowdsourced corpus into a high-quality dataset suitable to train and evaluate NLP systems requires, however, an aggregation method appropriate to the data and capable of achieving sufficient quality, something that simple majority voting typically cannot guarantee (Dawid and Skene, 1979; Hovy et al., 2013). What made it possible to extract such a dataset from the collected judgments was the recent development of a probabilistic method for aggregating coreference annotations called MPA (Paun et al., 2018b). MPA extracts silver labels from a coreference annotation and associates them with a probability, allowing for multiple labels in cases of ambiguity. As far as we know, ours is the first use of MPA to create a large-scale dataset. We show in the paper that MPA can be used to extract from the judgments a high quality coreference dataset that can be used to develop standard coreference resolvers, as well as to investigate disagreements on anaphora.

2 Background

2.1 Datasets for Anaphora/Coreference

Since the two CONLL shared tasks (Pradhan et al., 2012), ONTONOTES has become the dominant resource for anaphora resolution research (Fernandes et al., 2014; Björkelund and Kuhn, 2014; Martchat and Strube, 2015; Clark and Manning, 2015, 2016a,b; Lee et al., 2017, 2018). ONTONOTES contains documents in three languages: Arabic (300K tokens), Chinese (950K) and English (1.6M), from several genres but predominantly news. One frequently discussed limitation of ONTONOTES is the absence of singletons (De Marneffe et al., 2015; Chen et al., 2018), which makes it harder to train models for mention detection (Poesio et al., 2018). Another limitation is that expletives are not annotated. As a consequence, downstream applications such as machine translation (Guillou and Hardmeier, 2016) that require pronoun interpretation have to adopt various workarounds. Because of these two restrictions, ONTONOTES only has 195K markables, and a low markable density (0.12 markable/token).

A number of smaller corpora provide linguistically richer information (Poesio et al., 2016). Examples include ANCORA for Spanish (Recasens and Martí, 2010), TUBA-D/Z for German (Hinrichs et al., 2005), the Prague Dependency Treebank for Czech and English (Nedoluzhko et al., 2009), and ARRAU for English (Uryupina et al., To Appear). In ARRAU, for example, singletons and expletives are annotated as well, as are split antecedent plurals, generic coreference, discourse deixis, and bridging references. The ARRAU corpus is relatively small in terms of tokens (350K), but has a higher markable density than ONTONOTES (0.29 markable/token), so it has around 100K markables, half the number of ONTONOTES. ARRAU was recently used in the CRAC 2018 shared task (Poesio et al., 2018) to evaluate a number of anaphora resolution tasks.

The recently introduced PRECO corpus (Chen et al., 2018) is the largest existing coreference corpus, consisting of 35,000 documents for a total of 12.5M tokens and 3.8M markables, half of which are singletons. However, the corpus is not intended as a general purpose dataset as only the 3000 most common English words appear in the documents (the majority - 2/3 - of the documents are from Chinese high-school English tests). The corpus’s annotation scheme mainly follows the ONTONOTES guidelines, with a few important differences: singleton mentions and generic coreference are annotated, event anaphora is not, and predicative NPs are annotated as co-referring with their argument, as previously done in the MUC (Grishman and Sundheim, 1995; Chinchor, 1998) and ACE (Doddington et al., 2004) corpora. As one could expect, the corpus is relatively easy for coreference systems. The Peters et al. (2018) system trained and tested on PRECO achieves an av-

---

1 An example of predicative NP is 24 degrees in The temperature is 24 degrees. As discussed by van Deemter and Kibble (2000), annotating The temperature and 24 degrees as coreferent would result in nonsensical coreference chains for sentences like The temperature was 24 degrees but it is 27 degrees now. As a result, such markables were annotated as predicative in recent corpora. It’s not clear why we find a return to the old practice in PRECO.
verage CONLL score of 81.5%, whereas the same system trained and tested on ONTONOTES only achieves a score of 70.4%.

2.2 Crowdsourcing and GWAPs for NLP

A revolution in the way language annotation tasks are carried out was achieved by crowdsourcing (Howe, 2008; Snow et al., 2008). Crowdsourcing comes in many forms, including citizen science and microworking. A third approach is to use a game-with-a-purpose (GWAP) to aggregate data from non-expert players for collective decisions similar to those from an expert (von Ahn, 2006). The game-based approach to collecting language data is initially costly, but once a game is deployed it can continue to collect data with very little financial support, especially if there is an active community. GWAPs such as Phrase Detectives (Poesio et al., 2013), JeuxDesMots (Joubert and Lafourcade, 2008) and Zombie Lingo (Fort et al., 2014) have been used in NLP to collect data on specific linguistic features; broader platforms such as Wordrobe (Venhuizen et al., 2013) to gamify the entire text annotation pipeline.

Crowdsourcing is the most realistic approach to collect a large number of judgments about phenomena such as anaphora. Games in particular are the one type of crowdsourcing scalable to the goal of, for example, a 100M word corpus. So far, however, only small and medium scale resources for NLP have been created via crowdsourcing. For coreference we are only aware of two, both around 50K tokens in size (Chamberlain et al.; Guha et al., 2015). The Groningen Meaning Bank being collected through the Wordrobe platform (Bos et al., 2017) includes many more documents, but so far only very few interpretations have been obtained through the games (e.g., only around 4K judgments have been collected for anaphora).

2.3 Collecting Multiple Judgments

In most of the best known efforts at creating anaphoric corpora for English and other languages substantial disagreements between the coders were observed, but none of the resulting resources contains multiple anaphoric interpretations. Systematic analyses of the disagreements among coders observed in such annotation efforts were provided for ANCORA by Recasens et al. (2011) and for TUBA-D/Z by Versley (2008). The entire ONTONOTES corpus was double annotated, finding disagreements on around 20% of the markables, i.e., around 40,000 cases. An analysis of such disagreements can be found in (Pradhan et al., 2012), but ultimately only the result of adjudication was included in the corpus. Most of the PRECO corpus was doubly annotated and the results adjudicated, but only the result of adjudication is released.

We are aware of only two corpus annotation schemes which explicitly allowed the annotation of anaphoric ambiguity: ARRAU and the Potsdam Commentary Corpus (Krasavina and Chiarosc, 2007). Most of the ARRAU corpus was single-annotated by a highly experienced annotator, who was allowed to mark a variety of cases of ambiguity (Poesio and Artstein, 2005b). It is known, however, that such explicit marking of ambiguity is difficult (Poesio and Artstein, 2005b; Recasens et al., 2012), and indeed not many cases of ambiguity were marked in this way in ARRAU.

3 Collecting the Judgments

In this Section we discuss what type of judgments were collected, and how.

3.1 A gamified approach

The gamified platform Phrase Detectives² (Chamberlain et al., 2008; Poesio et al., 2013) was used to collect the judgments about anaphoric reference included in the corpus. Phrase Detectives is articulated around a number of tasks centered around the detective metaphor and uses scoring, progression and a variety of other mechanisms to make the activity enjoyable. In annotation mode (Name the Culprit), the participant provides an anaphoric judgment about a highlighted markable. If different participants enter different interpretations for a markable then each interpretation is presented to other participants in validation mode (Detectives Conference), in which the participants have to agree or disagree with the interpretation.

One of the key differences between Phrase Detectives and GWAPs such as those developed by von Ahn and his lab (von Ahn, 2006) is the much greater complexity of judgments required. Yet clearly we cannot expect participants to be experts about anaphora, or to be willing to read a manual explaining the annotation scheme, so all the training still has to be done while playing the game.

²http://www.phrasedetectives.com
Therefore, we developed a number of mechanisms that could help in this respect: giving suggestions and tips (global, contextual and FAQ), comparing decisions with the gold standard, and showing agreement with other players in Validation Mode. When participants begin to play they are shown training texts (in which the answer is known from a gold standard) and get feedback as to whether their decisions agree with the gold standard. Once the player has completed all training tasks they are given a user rating (the percentage of correct decisions out of the total number of training tasks).

As of 17th of March 2019, 60,822 individuals have participated in Phrase Detectives over ten years and using different platforms, providing over 4.26 million judgments, about half of which are included in the present release.

### 3.2 Types of Judgments

The judgments asked to the participants to Phrase Detectives follow a simplified version of the ARRAU annotation scheme, which is the result of extensive tests for intercoder agreement (Uryupina et al., To Appear). The participants are asked to make two basic distinctions: whether a markable is referring or not, and if referring, whether it is Discourse-Old (DO), i.e., it refers to an entity already mentioned (in which case the players were asked to indicate the latest mention of that entity), or Discourse-New (DN), i.e., it introduces a new entity in the discourse. Anaphoric reference marked include split antecedent anaphora, as in *John met Mary, and they went out for drinks*, where the antecedent for *they* is the set consisting of the separately introduced *John* and *Mary*. Two types of non-referring expressions were marked: expletives, as in *It’s five o’clock or There was a fireplace in the room*; and predicative NPs, as in *The temperature is 24 degrees*. In the case of predicative NPs, players were asked to mark the nearest mention of the entity that the predication applied to, following in this case the ONTONOTES approach instead of ARRAU’s.

The key difference between this corpus and any other existing corpus for anaphora / coreference with the exception of ARRAU is that the corpus was designed to collect information about disagreement. The main difference from ARRAU is that no attempt was made to ask players to identify ambiguity, as that has proven hard or impossible to do (Poesio and Artstein, 2005b). Instead of explicit (marking of) ambiguity, the developers relied on implicit ambiguity: that genuine ambiguity would emerge if enough players supplied judgments. All the judgments produced by the players were therefore stored, without attempting to choose among them at collection.

The differences between the four corpora being compared are summarized in Table 1, modelled on a similar table in (Chen et al., 2018). In the Phrase Detectives corpus predication and coreference are clearly distinguished, as in ONTONOTES and ARRAU but unlike in PRECO. Singletons are considered markables. Expletives and split antecedent plurals are marked, unlike in either ONTONOTES or PRECO. Most importantly, ambiguity of anaphoric interpretation (as in the example from the TRAINS corpus (Poesio and Artstein, 2005b)) is marked, but implicitly, i.e., by asking the judgment of at least 8 players per markable, as opposed to explicitly, as attempted in ARRAU (with little success).

### 3.3 Markable identification

Following standard practice in anaphoric annotation and GWAPS, the markables to be annotated were not identified by the participants themselves; instead, markable identification was carried out semi-automatically. Each document would first be processed by a pipeline combining off-the-shelf tools (sentence splitting and tokenization using the OpenNLP pipeline\(^3\) and parsing using the Berkeley Parser (Petrov and Klein, 2007)) and custom preprocessing and post-processing heuristic steps to correct the output. (See (Poesio et al., 2013) for more details about the pipeline and its performance.) Then one of the administrators would carry out a quick check of the document removing the most obvious mistakes before uploading it. After the document was uploaded, participants could report markable errors, which would then be corrected by hand.\(^4\)

### 4 The corpus

#### 4.1 Basic statistics

This second release of the Phrase Detectives corpus consists of a total of 542 documents contain-\(^3\)http://opennlp.apache.org

\(^4\)As participants report over 10,000 errors per year, it became quickly apparent that carrying out the corrections ourselves was unfeasible. In subsequent work, we developed a gamified approach to markable identification and correction centered around the TileAttack! GWAP (Madge et al., 2017).
Table 1: Comparison between the annotation schemes in ONTONOTES, PRECO, ARRAU and the present corpus

<table>
<thead>
<tr>
<th>Type</th>
<th>Example</th>
<th>ONTONOTES</th>
<th>PRECO</th>
<th>ARRAU</th>
<th>Present corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>predicative NPs</td>
<td>[John] is a teacher</td>
<td>Pred</td>
<td>Coref</td>
<td>Pred</td>
<td>Pred</td>
</tr>
<tr>
<td></td>
<td>[John, a teacher]</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>singletons</td>
<td>It's five o'clock</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>expletives</td>
<td>[Parents] are usually busy.</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>split antecedent plurals</td>
<td>[John] met [Mary] and they...</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>generic mentions</td>
<td>[Parents] should get involved</td>
<td>Only with</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>event anaphora</td>
<td>Sales [grew] 10%. This growth is exciting</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>ambiguity</td>
<td>Hook up [the engine] to [the boxcar] and send it to Avon</td>
<td>No</td>
<td>No</td>
<td>Explicit</td>
<td>Implicit</td>
</tr>
</tbody>
</table>

Table 2: Summary of the contents of the current release. The numbers in parentheses indicate the total number of markables that are non-singletons.

<table>
<thead>
<tr>
<th></th>
<th>Docs</th>
<th>Tokens</th>
<th>Markables</th>
</tr>
</thead>
</table>
| **PD
gold**         |      |        |           |
| Gutenberg            | 5    | 7536   | 1947 (1392) |
| Wikipedia            | 35   | 15287  | 3957 (1355) |
| GNOME                | 5    | 989    | 274 (96)   |
| Subtotal             | 45   | 23812  | 6178 (2843) |
| **PD
dilver**       |      |        |           |
| Gutenberg            | 145  | 158739 | 41989 (26364) |
| Wikipedia            | 350  | 218308 | 57678 (19444) |
| Other                | 2    | 7294   | 2126 (1339) |
| Subtotal             | 497  | 384341 | 101793 (47147) |
| **All**              | 542  | 408153 | 107971 (49990) |

Table 3: Percentage of markables with X distinct interpretations

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
</table>
| **PD
gold**         | 38.8 | 30.6 | 18.5 | 7.3 | 2.5 | 1.0 | 0.6 |
| **PD
dilver**       | 36.0 | 30.0 | 19.0 | 8.8 | 3.8 | 1.8 | 0.8 |

194480 markables. In other words, although the current release of the corpus is only about 25% of the CONLL corpus in terms of tokens, it is 55.5% of its size in terms of annotated markables, i.e., actual training / testing items.

4.2 Number of Judgments

In total, 2,235,664 judgments from 1958 players are included in the current release, of which 1,358,559 annotations and 867,844 validations. On average, 20.6 judgments were collected per markable: 12.6 annotations and 8 validations. In addition, around 10K expert judgments were collected for the gold portion of the corpus from two expert annotators. This compares with 600K estimated judgments for the entire ONTONOTES corpus, about 3 per markable (total number of annotators not known), and around 10M for PRECO, also 3 per markable, from about 80 annotators.

4.3 Disagreement: a preliminary analysis

The 'raw' statistics about disagreement in the corpus are shown in Table 3. In total, only 35.7% of the markables in the corpus (38,579) were assigned only one interpretation by the participants, whereas 64.3% received more than one interpretation. This figure would seem to suggest massive ambiguity, but we are not saying that 64.3% of markables in the corpus are ambiguous. As already pointed out e.g. in (Pradhan et al., 2012),

5PD
gold is the dataset released in 2016 as Phrase Detectives corpus, Release 1 (Chamberlain et al.).
there are a number of reasons for disagreements among coders / players apart from ambiguity. In the case of ONTONOTES, the causes for the 20,000 observed disagreements include:

- Ambiguity proper, i.e., unclear interpretation (‘Genuine Ambiguity’ in (Pradhan et al., 2012)) and/or disagreement on reference (31% of the disagreements in ONTONOTES, around 7% of all markables);
- Annotator error (another 25% of the cases of disagreement in ONTONOTES);
- Various limitations of the coding scheme: unclarity in the guidelines, inability to mark certain types of coreference e.g., between generics, etc. (36.5% of the cases of disagreement in ONTONOTES).
- Interface limitations (around 7.5% of the disagreements in ONTONOTES).

Some of the disagreements due to other causes—and in particular annotation errors—can be filtered through validation, i.e., by excluding those interpretations of a markable for which the validation score (annotations + agreements - disagreements) falls below a threshold. For example, if only interpretations with a validation score > 0 are considered, we find that 51,075 / 107,971 markables have at least two such interpretations, i.e., 47.3% of the total, which is considerably less than the 64.3% of markables with more than one interpretation, but it’s still a large number.

We will discuss a more sophisticated method for automatically identifying plausible interpretations, as well as the results of a preliminary hand-analysis of the disagreements in a few documents in our corpus, in Section 7.

5 Aggregation

5.1 Probabilistic Aggregation Methods

The models offer a rich framework of interpretation and can employ distinct prior and likelihood structures (pooled, unpoled, and partially poled) and a diverse set of effects (annotator ability, item difficulty, or a subtractive relationship between the two). However, most work on models of annotation assumes that the set of classes the annotators can choose from is fixed across the annotated items, which is not the case for anaphoric annotation. More specifically, in Phrase Detectives the participants can classify a markable as non-referring (expletive or predicative); as introducing a new discourse entity; or as discourse-old, in which case they link it to the most recent mention of its antecedent—and coreference chains are document-specific and not fixed in number (see Section 3.2 for more details on the annotation scheme). Recently, however, Paun et al. (2018b) developed a probabilistic model (MPA) able to aggregate such crowdsourced anaphoric annotations.

5.2 MPA

In MPA, the term label is used to refer to a specific interpretation provided by a player, and the term class to refer to general interpretation categories such as discourse old, discourse new, expletive, or predicative NP. Please note that under this formalism each label belongs to a class: the antecedent belongs to the discourse old category, while the other possible labels (e.g., discourse new) coincide with the classes they belong to. The model assumes a preprocessing step in which the markable-level annotations are transformed into a series of binary decisions with respect to each candidate label. MPA then models these (label-level) decisions as the result of the sensitivity (the true positive rate) and specificity (the true negative rate) of the annotators which it assumes are class dependent. This latter assumption allows inferring different levels of annotator ability for each class (thus capturing, for instance, the fact that whereas most participants are generally able to recognize discourse-new mentions, they are much less good at identifying correct antecedents).

5.3 Aggregating the game data

We use the MPA model as a component in a standard mention-pair framework to extract coreference clusters: 1) link each markable with the most likely label as identified by the model, and 2) follow the link structure to build the coreference
chains. We next evaluate both of these components against expert annotations.

Table 4 shows a per class evaluation of the aggregated interpretations from the PD_{gold} subset. The results indicate an overall better agreement with the expert annotations of MPA compared with a simple majority voting (MAJVOTE) baseline. This is because MAJVOTE makes the an implicit assumption that the annotators have equal expertise, which is not true in general even with data crowdsourced on microworking platforms, and even more so with data collected through GWAPs (Paun et al., 2018a).

After inferring the mention pairs, coreference chains can be extracted and their quality assessed using standard coreference metrics. Table 5 presents the evaluation against gold chains in PD_{gold}. We compare the chains produced from the mention pairs inferred by MPA and by MAJVOTE, and the chains produced by the STANFORD deterministic coreference system (Lee et al., 2011) (for which we switched off post-processing to output singleton clusters). The results indicate a far better quality of the chains produced using MPA over the alternative methods. Another interesting result is that even a simple MAJVOTE baseline based on crowdsourced annotations performed far better than the STANFORD system, underlining the advantage of crowdsourced annotations for coreference over automatically produced annotations.

### Table 4: A per class evaluation of aggregated interpretations against expert annotations.

<table>
<thead>
<tr>
<th>Method</th>
<th>Discourse old class</th>
<th>Discourse new class</th>
<th>Predicative NPs class</th>
<th>Expletives class</th>
<th>Avg. F1</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>P</td>
<td>R</td>
<td>F1</td>
</tr>
<tr>
<td>MAJVOTE</td>
<td>94.5</td>
<td>62.8</td>
<td>75.4</td>
<td>79.1</td>
<td>99.0</td>
<td>87.9</td>
</tr>
<tr>
<td>MPA</td>
<td>90.4</td>
<td>87.5</td>
<td>88.8</td>
<td>94.5</td>
<td>96.0</td>
<td>95.3</td>
</tr>
</tbody>
</table>

### Table 5: The quality of the coreference chains for the PD_{gold} subset.

<table>
<thead>
<tr>
<th>Method</th>
<th>MUC</th>
<th>BCUB</th>
<th>CEAPE</th>
<th>Avg. F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Singletons included</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAJVOTE</td>
<td>96.0</td>
<td>63.9</td>
<td>76.7</td>
<td>95.7</td>
</tr>
<tr>
<td>MPA</td>
<td>91.6</td>
<td>82.4</td>
<td>86.8</td>
<td>94.8</td>
</tr>
<tr>
<td>STANFORD</td>
<td>65.4</td>
<td>62.4</td>
<td>63.8</td>
<td>78.9</td>
</tr>
<tr>
<td>Singletons excluded</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAJVOTE</td>
<td>96.1</td>
<td>64.8</td>
<td>77.4</td>
<td>93.8</td>
</tr>
<tr>
<td>MPA</td>
<td>92.2</td>
<td>89.2</td>
<td>90.7</td>
<td>88.1</td>
</tr>
<tr>
<td>STANFORD</td>
<td>65.7</td>
<td>62.1</td>
<td>63.9</td>
<td>50.3</td>
</tr>
</tbody>
</table>

6 Using the corpus for coreference resolution

Some NLP researchers may question the usefulness of the information about disagreements for coreference resolution (or other NLP tasks). In this Section, we demonstrate that even those purely interested in CONLL-style coreference resolution can use the Phrase Detectives corpus aggregated with MPA as a dataset. We use PD_{silver} to train a coreference system able to simultaneously identify non-referring expression and build coreference chains (including singletons). As no other system of this type exists at the moment, we developed one ourselves.

6.1 Our system

The system trained and tested on the corpus is a cluster ranking system that does mention detection and coreference resolution jointly. The system uses the mention representation from the state-of-the-art (Lee et al., 2018) system, but replaces their mention-ranking model with a cluster ranking model. Our cluster ranking model forms clusters by going through the candidate mentions in their text order and adding them to the clusters, which take into consideration the relative importance of the mentions. An attention mechanism is used to assign mentions within the clusters salience scores, and the clusters are represented as the weighted sums of the mention representations. Separate classifiers are used to identify non-referring markables and singletons.
Table 6: The CoNLL scores for our systems trained on PD\textsubscript{silver} and tested on PD\textsubscript{gold}. * indicates the models trained on the simplified corpus.

<table>
<thead>
<tr>
<th>Singletons</th>
<th>Method</th>
<th>MUC P</th>
<th>R</th>
<th>F1</th>
<th>BCUB P</th>
<th>R</th>
<th>F1</th>
<th>CEAFE P</th>
<th>R</th>
<th>F1</th>
<th>Avg. F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Included</td>
<td>Our Model</td>
<td>79.3</td>
<td>72.5</td>
<td>75.7</td>
<td>72.1</td>
<td>69.3</td>
<td>70.7</td>
<td>70.5</td>
<td>73.2</td>
<td>71.8</td>
<td>72.7</td>
</tr>
<tr>
<td>Excluded</td>
<td>Our Model</td>
<td>79.3</td>
<td>72.5</td>
<td>75.7</td>
<td>58.3</td>
<td>52.4</td>
<td>55.2</td>
<td>58.3</td>
<td>49.5</td>
<td>53.5</td>
<td>61.5</td>
</tr>
<tr>
<td></td>
<td>Our Model*</td>
<td>77.8</td>
<td>71.8</td>
<td>74.6</td>
<td>55.4</td>
<td>53.7</td>
<td>54.6</td>
<td>56.2</td>
<td>49.0</td>
<td>52.4</td>
<td>60.5</td>
</tr>
<tr>
<td></td>
<td>Lee et al. (2018)*</td>
<td>80.8</td>
<td>66.1</td>
<td>72.7</td>
<td>63.3</td>
<td>45.1</td>
<td>52.7</td>
<td>56.7</td>
<td>44.7</td>
<td>50.0</td>
<td>58.5</td>
</tr>
</tbody>
</table>

Table 7: Non-referring scores for our model

<table>
<thead>
<tr>
<th>Type</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-referring</td>
<td>55.2</td>
<td>54.0</td>
<td>54.6</td>
</tr>
<tr>
<td>Expletives</td>
<td>62.3</td>
<td>86.0</td>
<td>72.3</td>
</tr>
<tr>
<td>Predicative NPs</td>
<td>49.7</td>
<td>47.7</td>
<td>48.7</td>
</tr>
</tbody>
</table>

6.2 Evaluation Methodology

We randomly chose 1/20 of PD\textsubscript{silver} as a development set and use the rest as the training set; PD\textsubscript{gold} was used as test set.

To get a baseline, we compare the results of our system on a simplified version of the corpus without singletons and expletives with those obtained by the current state-of-the-art system on ONTONOTES, Lee et al. (2018) trained and tested on the same data.

6.3 Results

Table 6 shows the results of both systems on the simplified corpus. Our cluster ranking system achieved an average CoNLL score of 60.5%, outperforming the Lee et al. (2018) system by 2 percentage points. Note that the Lee et al. (2018) system achieved a higher score on the CONLL data, which suggests that the present corpus is different from that dataset.

In the same Table, we also report the results obtained by training our system on the full corpus including both non-referring expressions and singletons. This version of system achieves an average CONLL score of 72.7%. We will note that although this score is on system mentions, it is very close to the score (74.3%) achieved by the Stanford deterministic system evaluated with gold mentions (see Table 5 in Section 5). Also, this model trained on the full corpus including singletons achieves a gain of 1 percentage point when compared with the model trained on the simplified corpus even when evaluated in a singleton excluded setting. This indicates that the availability of the singletons is also helpful for resolving non-singleton clusters. In total, this model achieved a CONLL score on singletons excluded of 3 percentage points 3% better than our baseline.

Regarding the task of identifying non-referring mentions, our model achieved a F1 score of 54.6% (see Table 7). The scores of the system on distinct types of non-referring expressions is presented in the following two rows of Table 7. Our model achieved a higher F1 score of 72.3% on expletives, and a lower score (48.7%) on predicative NPs.

Overall, these results—the first results on system mentions for PD\textsubscript{gold}—suggest that the silver corpus is sufficient to train a state-of-the-art system and achieve a reasonably good performance. Also, that training a model on a corpus enhanced by singletons and non-referring markables results in a better CONLL score when compared with a model trained on the simplified corpus.

7 Disagreements, revisited

In the previous Section we showed that MPA can be used to extract a silver standard out of the annotations that is suitable to train a CONLL-style coreference resolver or an extended coreference resolver also attempting identification of singletons and non-referring expressions. The key property of the corpus however is the information it provides about disagreements. The second useful contribution of MPA is that it can be used to get an assessment of the ambiguity of markables which is more refined than that discussed in Section 4.3. For each markable, MPA assigns a probability to each interpretation. Given that the model does not assume the existence of a ‘gold’, there are three possible cases for each markable: either only one interpretation has a probability above a...
certain threshold—say, 0.5; or more than one interpretation is above that threshold; or none is. This assessment of ambiguity according to MPA is summarized in Table 8.

This assessment appears to suggest a similar prevalence of ambiguity in our corpus than found in ONTONOTES in the already mentioned analysis by Pradhan et al. (2012). In order to verify this, two experts hand-analyzed 2 documents in PDgold containing a total of 900 markables: Little Red Cap (LRC) and Robber Bridegroom (RG). Given that each markable has on average 20 interpretations, and that player errors are frequent (there is at least one player error for almost every markable) it wasn’t possible to use the same categories as Pradhan et al. Instead, we simply attempted to assign markables to one of the categories: Genuine ambiguity (GA), Interface or Coding Scheme Problem (ICP), Other (O). The results are summarized in Table 9. The Table has one row per document. The first column lists the total number of markables in a document; the second (Dis) the percentage of markables on which there is disagreement; the third (GA) the percentage of the total number of markables which are cases of genuine ambiguity; and the fourth (ICP) the percentage which are cases of Interface or Coding Scheme Problem. As we can see from the Table, 9% of the total number of markables in these documents (80 out of 865) are genuinely ambiguous, i.e., that 12.6% of the disagreements (80 out of 633) are cases of genuine ambiguity. These are only preliminary figures, and we suspect that the ultimate figures on the prevalence of ambiguity are going to be much higher, given that Recasens et al. (2012) report that 12-15% of coreference relations in their corpus are cases of quasi-coreference, and that Poesio and Artstein (2005a) report a figure of 42.6% once ambiguity on discourse deictic reference are taken into account.

We next checked the extent to which MPA can correctly predict genuine ambiguity. The results suggest that MPA is good at removing spurious ambiguity, but as a predictor of ambiguity it only has a recall of around 20% and a precision of slightly under 50%. Improving these results is one of the objectives of our current research.

8 Conclusions

We presented a novel resource for anaphora that, because of its annotation scheme and size, at the very least should be useful to those in the community interested in developing systems able to perform a more comprehensive form of anaphora resolution, including for instance expletive detection and split antecedent resolution. The key property of this new resource however is that it provides a large number of judgments about each anaphoric expression, thus enabling the development of systems that do not make the assumption that a ‘gold standard’ exists, an assumption questioned by all studies associated with the creation of the current resources for the task. The dataset is also to our knowledge the first solid evidence that the games-with-a-purpose approach can be successfully deployed to obtain substantial resources for NLP.

The corpus is freely available from the Linguistic Data Consortium and from http://www.dali-ambiguity.org. It is distributed in three formats: an XML format including all the judgments, suitable for analysis of disagreements and/or the development of systems taking disagreement into account; and in CONLL and CRAC18 format, with only the gold annotation or the silver label extracted, for those interested in using the corpus as an alternative resource for developing coreference systems only.

Acknowledgements

This research was supported by the DALI project, funded by the European Research Council (ERC), Grant agreement ID: 695662.
References


Jon Chamberlain, Massimo Poesio, and Udo Kruschwitz. phrase detectives corpus.


Kian Kenyon-Dean, Eisha Ahmed, Jeremy Georges-Filteau, Christopher Glaz, Barleen Kaur, Auguste Lalande, Shrutí Bhanderi, Robert Belfer, Nirmał Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths. 2018. Sentiment analysis:
it’s complicated! In Proceedings of NAACL, pages 1886–1895. ACL.


Mathieu Lafourcade, Alain Joubert, and Nathalie Le Brun. 2015. Games with a Purpose (GWAPs). Wiley.


Barbara Plank, Dirk Hovy, and Anders Sogaard. 2014. Linguistically debatable or just plain wrong? In Proceedings of EACL.


Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa J. Rodriguez, and Massimo Poesio. To Appear. Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU corpus. *Journal of Natural Language Engineering*.


