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 word-senses (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; Martschat 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.[1] 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.

erage 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 Chiarcos, 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 online platform *Phrase Detectives*[2] (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 (the possible judgments according to the annotation scheme are discussed next). 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.

[2] http://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).

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

Table 1: Comparison between the annotation schemes in ONTONOTES, PRECO, ARRAU and the present corpus

		Docs	Tokens	Markables
PD_{gold}	Gutenberg	5	7536	1947 (1392)
	Wikipedia	35	15287	3957 (1355)
	GNOME	5	989	274 (96)
	Subtotal	45	23812	6178 (2843)
$\text{PD}_{\text{silver}}$	Gutenberg	145	158739	41989 (26364)
	Wikipedia	350	218308	57678 (19444)
	Other	2	7294	2126 (1339)
	Subtotal	497	384341	101793 (47147)
All	Total	542	408153	107971 (49990)

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

	1	2	3	4	5	6	7
PD_{gold}	38.8	30.6	18.5	7.3	2.5	1.0	0.6
$\text{PD}_{\text{silver}}$	36.0	30.0	19.0	8.8	3.8	1.8	0.8

Table 3: Percentage of markables with X distinct interpretations

ing 408K tokens and 108K markables from two main genres: Wikipedia articles and fiction from the Gutenberg collection. This corpus is divided in two subsets. The subset we refer to as $\text{PD}_{\text{silver}}$ consists of 497 documents, for a total of 384K tokens and 101K markables, whose annotation was completed–i.e. 8 judgments per markable were collected, and 4 validations per interpretation–as of 12th of October 2018. In these documents, an aggregated ('silver') label obtained through MPA (see next Section) is also provided. 45 additional documents were also gold-annotated by two experts annotators. We refer to the subset of the corpus for which both gold and silver annotations are available as PD_{gold}, as it is intended to be used as test set.[5] The gold subset consists of a total of 23K tokens and 6K markables. The contents of the corpus are summarized in Table 2.

By comparison, the English corpus used for the CONLL 2011 and 2012 shared tasks consists of 3493 documents, for a total of 1.6M tokens and 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),

[5] $\text{PD}_{\text{gold}}$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 data collected via *Phrase Detectives* require an aggregation method to help choose between the different interpretations provided by the players. Simple heuristics such as majority voting are known to underperform compared to probabilistic models of annotation (Whitehill et al., 2009; Raykar et al., 2010; Quoc Viet Hung et al., 2013; Sheshadri and Lease, 2013; Hovy et al., 2013; Passonneau and Carpenter, 2014; Paun et al., 2018a). The models offer a rich framework of interpretation and can employ distinct prior and likelihood structures (pooled, unpooled, and partially pooled) 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 antecedents belong 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

Method	Discouse old class			Discourse new class			Predicative NPs class			Expletives class			Avg. F1	Accuracy
	P	R	F1	P	R	F1	P	R	F1	P	R	F1		
MAJVOTE	**94.5**	62.8	75.4	79.1	**99.0**	87.9	53.9	9.7	16.4	**97.2**	71.4	82.4	65.5	82.9
MPA	90.4	**87.3**	**88.8**	**94.5**	96.0	**95.3**	**64.0**	**72.4**	**68.0**	94.1	**98.0**	**96.0**	**87.0**	**92.2**

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

	Method	MUC			BCUB			CEAFE			Avg. F1
		P	R	F1	P	R	F1	P	R	F1	
Singletons included	MAJVOTE	**96.0**	63.9	76.7	**95.7**	78.7	86.4	77.1	**94.9**	85.1	82.7
	MPA	91.6	**82.4**	**86.8**	94.8	**87.8**	**91.2**	**92.4**	93.8	**93.1**	**90.3**
	STANFORD	65.4	62.4	63.8	78.9	76.1	77.5	78.4	85.2	81.7	74.3
Singletons excluded	MAJVOTE	**96.1**	64.8	77.4	**93.8**	45.0	60.8	66.3	48.5	56.1	64.8
	MPA	92.2	**89.2**	**90.7**	88.1	**77.8**	**82.6**	**79.5**	**80.2**	**79.8**	**84.4**
	STANFORD	65.7	62.1	63.9	50.3	42.5	46.1	42.7	49.8	46.0	52.0

Table 5: The quality of the coreference chains for the PD$_{\text{gold}}$ subset.

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$_{\text{gold}}$ 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$_{\text{gold}}$. 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.

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$_{\text{silver}}$ 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 corefence 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.

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

Table 6: The CoNLL scores for our systems trained on $\text{PD}_{\text{silver}}$ and tested on PD_{gold}. * indicates the models trained on the simplified corpus.

	P	R	F1
Non-referring	55.2	54.0	54.6
Expletives	62.3	86.0	72.3
Predicative NPs	49.7	47.7	48.7

Table 7: Non-referring scores for our model

6.2 Evaluation Methodology

We randomly chose 1/20 of $\text{PD}_{\text{silver}}$ as a development set and use the rest as the training set; PD_{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%.[6] 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_{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

[6] The Extended Coreference Scorer developed for the 2018 CRAC shared task (Poesio et al., 2018) was used to evaluate coreference chains on a corpus using singletons and to assess non-referring expressions identification.

	None	One	Two	Zero or more
PD_{gold}	2.3%	93.4%	4.3%	6.6%
$\text{PD}_{\text{silver}}$	3.5%	94%	2.4%	5.9%

Table 8: Ambiguity in the corpus according to MPA

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 PD_{gold} 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.

	Total	Dis	GA	ICP
LRC	401	79.1%	7% (28)	7.7% (31)
RG	464	68.3%	11.2% (52)	12.9% (60)
Average		73.7%	9.1%	10.3%

Table 9: Analysis of disagreements in two corpus documents

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

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

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Anders Björkelund and Jonas Kuhn. 2014. Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. In *Proceedings of the 52nd ACL*, volume 1, pages 47–57.

Johan Bos, Valerio Basile, Kilian Evang, Noortje J. Venhuizen, and Johannes Bjerva. 2017. The groningen meaning bank. In N. Ide and J. Pustejovsky, editors, *The Handbook of Linguistic Annotation*, chapter 18, pages 463–496. Springer.

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

Jon Chamberlain, Massimo Poesio, and Udo Kruschwitz. 2008. Phrase Detectives: A web-based collaborative annotation game. In *Proceedings of I-Semantics 2008*.

Hong Chen, Zhenhua Fan, Hao Lu, Alan Yuille, and Shu Rong. 2018. PreCo: A large-scale dataset in preschool vocabulary for coreference resolution. In *Proceedings of EMNLP*, pages 172–181, Brussels, Belgium.

Nancy Chinchor. 1998. Overview of MUC-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

Kevin Clark and Christopher D. Manning. 2015. Entity-centric coreference resolution with model stacking. In *Proceedings of the ACL*.

Kevin Clark and Christopher D. Manning. 2016a. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of EMNLP*.

Kevin Clark and Christopher D. Manning. 2016b. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the ACL*.

Daniel Clery. 2011. Galaxy evolution. Galaxy Zoo volunteers share pain and glory of research. *Science*, 333(6039):173–5.

Seth Cooper, Firsas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popovic, and the Foldit Players. 2010. Predicting protein structures with a multiplayer online game. *Nature*, 466:756–760.

Alexander P. Dawid and Allan M. Skene. 1979. Maximum likelihood estimation of observer errorrates using the EM algorithm. *Applied Statistics*, 28(1):20–28.

Marie-Catherine De Marneffe, Marta Recasens, and Christopher Potts. 2015. Modeling the lifespan of discourse entities with application to coreference resolution. *Journal of Artificial Intelligence Research*, 52(1):445–475.

Kees van Deemter and Rodger Kibble. 2000. On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics*, 26(4):629–637. Squib.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Proceedings of LREC*, volume 2, page 1.

Eraldo R. Fernandes, Cícero N. dos Santos, and Ruy L. Milidiú. 2014. Latent trees for coreference resolution. *Computational Linguistics*, 40(4):801–835.

Karen Fort, Bruno Guillaume, and H. Chastant. 2014. Creating Zombilingo, a game with a purpose for dependency syntax annotation. In *Proceedings of the 1st International Workshop on Gamification for Information Retrieval (GamifIR'14)*, pages 2–6. ACM.

Ralph Grishman and Beth Sundheim. 1995. Design of the muc-6 evaluation. In *Proceedings of the 6th Conference on Message Understanding*, pages 1–11. Association for Computational Linguistics.

Anupam Guha, Mohit Iyyer, Danny Bouman, and Jordan Boyd-Graber. 2015. Removing the training wheels: A coreference dataset that entertains humans and challenges computers. In *Proceedings of NAACL*.

Liane Guillou and Christian Hardmeier. 2016. Protest: A test suite for evaluating pronouns in machine translation. In *Proceedings of LREC*, Paris, France.

Erhard W. Hinrichs, Sandra übler, and Karin Naumann. 2005. A unified representation for morphological, syntactic, semantic and referential annotations. In *Proc. of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, Ann Arbor, Michigan.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of NAACL*, pages 1120–1130.

Jeff Howe. 2008. *Crowdsourcing: Why the power of the crowd is driving the future of business*. Crown Publishing Group.

Alan Joubert and Mathieu Lafourcade. 2008. Jeuxdemots : Un prototype ludique pour l'émergence de relations entre termes. In *Proceedings of JADT*.

Kian Kenyon-Dean, Eisha Ahmed, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths. 2018. Sentiment analysis:

it's complicated! In *Proceedings of NAACL*, pages 1886–1895. ACL.

Olga Krasavina and Christian Chiarcos. 2007. The Potsdam coreference scheme. In *Proceedings of the 1st Linguistic Annotation Workshop*, pages 156–163.

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

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of CONLL: Shared Task*, pages 28–34, Stroudsburg, PA, USA.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of EMNLP*.

Kenton Lee, Luheng He, and Luke S. Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of ACL*.

Chris Madge, Jon Chamberlain, Udo Kruschwitz, and Massimo Poesio. 2017. Experiment-driven development of a gwap for marking segments in text. In *Proceedings of CHI PLAY*, Amsterdam.

Sebastian Martschat and Michael Strube. 2015. Latent structures for coreference resolution. *Transactions of the Association for Computational Linguistics*, 3:405–418.

A. Nedoluzhko, J. Mirokvský, and P. Pajas. 2009. The coding scheme for annotating extended nominal coreference and bridging anaphora in the prague dependency treebank. In *Proceedings of the Linguistic Annotation Workshop*, pages 108–111.

Rebecca J. Passonneau, Vikas Bhardwaj, Ansaf Salleb-Aouissi, and Nancy Ide. 2012. Multiplicity and word sense: evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation*, 46(2):219–252.

Rebecca J. Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018a. Comparing bayesian models of annotation. *Transactions of the Association for Computational Linguistics*.

Silviu Paun, Jon Chamberlain, Udo Kruschwitz, Juntao Yu, and Massimo Poesio. 2018b. A probabilistic annotation model for crowdsourcing coreference. In *Proceedings of EMNLP*, pages 1926–1937, Brussels, Belgium. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke S. Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL*.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Proceedings of HLT-NAACL*.

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

Massimo Poesio and Ron Artstein. 2005a. Annotating (anaphoric) ambiguity. In *Proc. of the Corpus Linguistics Conference*, Birmingham.

Massimo Poesio and Ron Artstein. 2005b. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation*, pages 76–83.

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi. 2013. Phrase Detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Interactive Intelligent Systems*, 3(1):1–44.

Massimo Poesio, Yulia Grishina, Varada Kolhatkar, Nafise Moosavi, Ina Roesiger, Adam Roussel, Fabian Simonjetz, Alexandra Uma, Olga Uryupina, Juntao Yu, and Heike Zinsmeister. 2018. Anaphora resolution with the arrau corpus. In *Proc. of the NAACL Worskhop on Computational Models of Reference, Anaphora and Coreference (CRAC)*, pages 11–22, New Orleans.

Massimo Poesio, Sameer Pradhan, Marta Recasens, Kepa Rodriguez, and Yannick Versley. 2016. Annotated corpora and annotation tools. In M. Poesio, R. Stuckardt, and Y. Versley, editors, *Anaphora Resolution: Algorithms, Resources and Applications*, chapter 4. Springer.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL 2012)*, Jeju, Korea.

Nguyen Quoc Viet Hung, Nguyen Thanh Tam, Lam Ngoc Tran, and Karl Aberer. 2013. An evaluation of aggregation techniques in crowdsourcing. In *Web Information Systems Engineering – WISE 2013*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322.

Marta Recasens, Ed Hovy, and M. Antònia Martí. 2011. Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121(6):1138–1152.

Marta Recasens and M. Antònia Martí. 2010. AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation*, 44(4):315–345.

Marta Recasens, M. Antonia Martí, and Constantin Orasan. 2012. Annotating near-identity from coreference disagreements. In *Proceedings of LREC*.

Aashish Sheshadri and Matthew Lease. 2013. SQUARE: A benchmark for research on computing crowd consensus. In *Proceedings of the 1st AAAI Conference on Human Computation (HCOMP)*, pages 156–164.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast - but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP*, pages 254–263.

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*.

Noortje Venhuizen, Valerio Basile, Kilian Evang, and Johan Bos. 2013. Gamification for word sense labeling. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS'13)*.

Yannick Versley. 2008. Vagueness and referential ambiguity in a large-scale annotated corpus. *Research on Language and Computation*, 6:333–353.

Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier R. Movellan, and Paul L. Ruvolo. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems 22*, pages 2035–2043. Curran Associates, Inc.

Billy T. M. Wong and Sophia Y. M. Lee. 2013. Annotating legitimate disagreement in corpus construction. In *Proceedings of IJCNLP*, pages 51–57, Nagoya, Japan.