Aggregating Crowdsourced and Automatic Judgments to Scale Up a Corpus of Anaphoric Reference for Fiction and Wikipedia Texts

Juntao Yu[1], **Silviu Paun**[2]*, **Maris Camilleri**[3], **Paloma Carretero Garcia**[3], **Jon Chamberlain**[1], **Udo Kruschwitz** [4] and **Massimo Poesio**[3]
[1]University of Essex, UK; [2]Amazon Research, Romania;
[3]Queen Mary University of London, UK; [4]University of Regensburg, Germany.
j.yu@essex.ac.uk; silviupn@amazon.com; mcamil@essex.ac.uk;pcarre@essex.ac.uk;
jchamb@essex.ac.uk; udo.kruschwitz@ur.de; m.poesio@qmul.ac.uk;

Abstract

Although several datasets annotated for anaphoric reference / coreference exist, even the largest such datasets have limitations in term of size, range of domains, coverage of anaphoric phenomena, and size of documents included. Yet, the approaches proposed to scale up anaphoric annotation haven't so far resulted in datasets overcoming these limitations. In this paper, we introduce a new release of a corpus for anaphoric reference labelled via a game-with-a-purpose. This new release[1] is comparable in size to the largest existing corpora for anaphoric reference due in part to substantial activity by the players, in part thanks to the use of a new resolve-and-aggregate paradigm to 'complete' markable annotations through the combination of an anaphoric resolver and an aggregation method for anaphoric reference. The proposed method could be adopted to greatly speed up annotation time in other projects involving games-with-a-purpose. In addition, the corpus covers genres for which no comparable size datasets exist (Fiction and Wikipedia); it covers singletons and non-referring expressions; and it includes a substantial number of long documents ($> 2K$ in length).

1 Introduction

Many resources annotated for anaphoric reference / coreference exist; but even the largest such datasets, such as ONTONOTES (Pradhan et al., 2012), have limitations. The largest resources are still medium scale (e.g., ONTONOTES (Pradhan et al., 2012) is 1.5M tokens, as is CRAFT (Cohen et al., 2017)). They only cover a limited range of domains, primarily news (as in ONTONOTES) and scientific articles (as in CRAFT), and models trained on these datasets have been shown not to generalize well to other domains (Xia and Durme, 2021).[2] The range of anaphoric phenomena covered is also narrow (Poesio et al., 2016).

Several proposals have been made to scale up anaphoric annotation in terms of size, range of domains, and phenomena covered proposed, including automatic data augmentation (Emami et al., 2019; Gessler et al., 2020; Aloraini and Poesio, 2021), and crowdsourcing combined with active learning (Laws et al., 2012; Li et al., 2020; Yuan et al., 2022) or through Games-With-A-Purpose (Chamberlain et al., 2008; Hladká et al., 2009; Bos et al., 2017; Kicikoglu et al., 2019). However, the largest existing anaphoric corpora created using Games-With-A-Purpose (e.g., (Poesio et al., 2019)) are still smaller than the largest resources created with traditional methods, and the corpora created using data augmentation techniques are focused on specific aspects of anaphoric reference. In order to use such approaches to create resources of the required scale in terms of size, variety and range of phenomena covered novel methods are required.

The first contribution of this paper is the *Phrase Detectives* 3.0 corpus of anaphoric reference annotated using a Game-With-A-Purpose. This corpus has a comparable size in tokens (1.37M) to the ONTONOTES corpus (Pradhan et al., 2012), but twice the number of markables[3]. Its annotation scheme also covers singletons and non-referring expressions; it is focused on two genres - fiction and Wikipedia articles - not covered in ONTONOTES, and for which only much smaller datasets exist; and it includes a range of documents ranging from short to fairly long ($14K$ tokens) enabling research on NLP in long documents (Beltagy et al., 2020). While ONTONOTES will remain a fundamental resource for the field in terms of size and languages it covers, we therefore hope that *Phrase Detec-*

*Work was done prior to joining Amazon research.

[1]The corpus is available at https://github.com/dali-ambiguity/Phrase-Detectives-Corpus-3.0

[2]The largest existing corpus for English, the 10M words PRECO (Chen et al., 2018), consists of language learning texts, but the models trained on this genre have proven to have even worse performance on other domains.

[3]The number of non-singleton markables is similar to that of ONTONOTES

tives 3.0 will complement ONTONOTES, providing a comparable amount of data in genres so far less covered, while at the same time covering aspects of anaphoric interpretation not covered there, such as singletons and non-referring expressions.

The second contribution of the paper is a new iterative resolve-and-aggregate approach developed to 'complete' the corpus by combining crowdsourcing with automatic annotation. Only about 70% of documents in the corpus were completely annotated by the players. The proposed method (i) uses an anaphoric resolver to automatically annotate all mentions, including the few still unannotated; (ii) aggregates the resulting judgments using a probabilistic aggregation method for anaphora, and (iii) uses the resulting expanded dataset to retrain the anaphoric resolver. We show that the resolve-and-aggregate method results in models with higher accuracy than models trained using only the completely annotated data, or the full corpus not completed using the method.

2 Background

Anaphorically annotated corpora A number of anaphorically annotated datasets now exist, covering a number of languages (Hinrichs et al., 2005; Hendrickx et al., 2008; Recasens and Martí, 2010; Pradhan et al., 2012; Landragin, 2016; Nedoluzhko et al., 2016; Cohen et al., 2017; Chen et al., 2018; Bamman et al., 2020; Uryupina et al., 2020; Zeldes, 2020) and turning anaphora / coreference in a very active area of research (Pradhan et al., 2012; Fernandes et al., 2014; Wiseman et al., 2015; Lee et al., 2017, 2018; Yu et al., 2020; Joshi et al., 2020). However, only a few of these are genuinely large in terms of markables (Pradhan et al., 2012; Cohen et al., 2017), and most are focused on news. Corpora of comparable size exist for scientific articles (e.g., CRAFT (Cohen et al., 2017)), and substantially smaller ones for fiction (e.g., LitBank (Bamman et al., 2020) and *Phrase Detectives* 2 (Poesio et al., 2019)), and Wikipedia (e.g., WikiCoref (Ghaddar and Langlais, 2016) or again *Phrase Detectives* 2 (Poesio et al., 2019)). But important genres such as dialogue are barely covered (Muzerelle et al., 2014; Yu et al., 2022a). There is evidence that this concentration on a single genre, and on ONTONOTES in particular, results in model that do not generalize well (Xia and Durme, 2021).

Existing resources are also limited in terms of coverage. Most recent datasets are based on general purpose annotation schemes with a clear linguistic foundation, but especially the largest ones focus on the simplest cases of anaphora / coreference (e.g., singletons and non-referring expressions are not annotated in ONTONOTES). And the documents included in existing corpora tend to be short, with the exception of CRAFT: e.g., average document length is 329 in PRECO, 467 in ONTONOTES, 630 in ARRAU, and 753 in *Phrase Detectives*.

Scaling up anaphoric annotation One approach to scale up anaphoric reference annotation is using fully automatic methods to either annotate a dataset, such as AMALGUM (Gessler et al., 2020), or create a benchmark from scratch, such as KNOWREF (Emami et al., 2019). While entirely automatic annotation may result in datasets of arbitrarily large size, such annotations cannot expand current models' coverage to aspects of anaphoric reference do not already handle well. And creating from scratch large-scale benchmarks for specific issues hasn't so far been shown to result in datasets reflecting the variety and richness of real texts.

Crowdsourcing has emerged as the dominant paradigm for annotation in NLP (Snow et al., 2008; Poesio et al., 2017) because of its reduced costs and increased speed in comparison with traditional annotation. But the costs for really large-scale annotation are still prohibitive even for crowdsourcing (Poesio et al., 2013, 2017). To address this issue, a number of approaches have been developed to optimize the use of crowdsourcing for coreference annotation. In particular, active learning has been used to reduce the amount of annotation work needed (Laws et al., 2012; Li et al., 2020; Yuan et al., 2022). Another issue is that anaphoric reference is a complex type of annotation whose most complex aspects require special quality control typically not available with microtask crowdsourcing.

Games-With-A-Purpose A form of crowdsourcing which has been widely used to address the issues of cost and quality is Games-With-A-Purpose (GWAP) (von Ahn, 2006; Cooper et al., 2010; Lafourcade et al., 2015). Games-With-A-Purpose is the version of crowdsourcing where labelling is created through a game, so that the reward for the workers is in terms of enjoyment rather than financial–were proposed as a solution for large-scale data labelling. A number of GWAPs were therefore developed for NLP, including *Jeux de Mots* (Lafourcade, 2007; Joubert et al., 2018),

Phrase Detectives (Chamberlain et al., 2008; Poesio et al., 2013), *OntoGalaxy* (Krause et al., 2010), the *Wordrobe* platform (Basile et al., 2012), *Dr Detective* (Dumitrache et al., 2013), *Zombilingo* (Fort et al., 2014), *TileAttack!* (Madge et al., 2017), *Wormingo* (Kicikoglu et al., 2019), *Name That Language!* (Cieri et al., 2021) or *High School Superhero* (Bonetti and Tonelli, 2021). GWAPs for coreference include *Phrase Detectives* (Chamberlain et al., 2008; Poesio et al., 2013), the *Pointers* game in *WordRobe* (Bos et al., 2017) and *Wormingo* (Kicikoglu et al., 2019), all deployed, and *PlayCoref* (Hladká et al., 2009), proposed but not tested.

However, whereas truly successful GWAPs such as FOLDIT have been developed in other areas of science (Cooper et al., 2010), even the most successful GWAPs for NLP only collected moderate amounts of data (Poesio et al., 2019; Joubert et al., 2018). In part, this is because the games used to actually collect NLP labels aren't very entertaining, leading to efforts to develop engaging designs such as (Jurgens and Navigli, 2014; Dziedzic and Włodarczyk, 2017; Madge et al., 2019).

An interesting solution to this issue was proposed although not fully developed for *Wordrobe* (Bos et al., 2017). This solution is a hybrid between automatic annotation and crowdsourcing: a combination of crowd and automatically computed judgments is aggregated to ensure that every item has at least one label. This solution wasn't properly tested in *Wordrobe*, which only collected very few judgments and for a small corpus; and anyway the approach followed could not be applied to anaphora/coreference, due to the lack of a suitable aggregation mechanism for anaphora/coreference. In this paper we present the first true test of the idea by proposing a method for aggregating crowd and automatic judgments inspired by this idea, but using an aggregation method for anaphora, and truly tested on a dataset containing a very large number of anaphoric judgments.

3 Phrase Detectives

The human judgments used in our corpus were collected using the *Phrase Detectives* Game-With-A-Purpose (Chamberlain et al., 2008; Poesio et al., 2013; Chamberlain, 2016; Poesio et al., 2019), designed to collect multiple judgments about anaphoric reference.

Game design *Phrase Detectives* doesn't follow the design of some of the original von Ahn games (von Ahn and Dabbish, 2008), in that it is a one-person game, and not timed; both competition and timing were found to have orthogonal effects on the quality of the annotation (Chamberlain, 2016). Points are used as the main incentive, with weekly and monthly boards being displayed.

Players play two different games: one aiming at labelling new data, the other at validating judgments expressed by the other players. In the annotation game, *Name the Culprit*, the player 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 the validation game, *Detectives Conference*, in which the participants have to agree or disagree with the interpretation.

Every item is annotated by at least 8 players (20 on average), and each distinct interpretation is validated by at least four players. Players get points for each label they produce, but especially when their interpretation is agreed upon by other players, thus rewarding accuracy. Initially, players play against gold data, and are periodically evaluated against the gold; when they achieve a sufficient level of accuracy, they start seeing incompletely annotated data. Extensive analyses of the data suggest that although there is a great number of noisy judgments, this simple training and validation method delivers extremely accurate aggregated labels (Poesio et al., 2013; Chamberlain, 2016; Poesio et al., 2019)

Annotation Scheme The annotation scheme used in *Phrase Detectives* is a simplified version of the ARRAU annotation scheme (Uryupina et al., 2020), covering all the main aspects of anaphoric annotation, including the distinction between referring and non-referring expressions (all noun phrases are annotated as either referring or non-referring, with two types of non-referring expressions being annotated: expletives and predicative NPs); the distinction between discourse-new and discourse-old referring expressions (Prince, 1992); and the annotation of all types of identity reference (including split antecedent plural anaphora). Only the most complex types of anaphoric reference (bridging references and discourse deixis) are not annotated. The main differences between the annotation scheme used in *Phrase Detectives* and those used in ARRAU, ONTONOTES, and PRECO are summarized in Table 1, modelled on a similar

Type	Example	ONTONOTES	PRECO	ARRAU	*Phrase Detectives*
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 *Phrase Detectives*.

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.

Possibly the most distinctive feature of the annotation scheme is that disagreements among annotators are preserved, encoding a form of implicit ambiguity as opposed to the explicit ambiguity annotated in ARRAU.

The DEV and TEST subsets of the corpus (see next Section) have been annotated according to the full ARRAU scheme.

Preliminary player statistics At the time of writing (11th October, 2022), 61,391 players have registered on *Phrase Detectives*, of which more than 4,000 demonstrated sufficient linguistic understanding that they were graduated to allowed to provide judgments on partially labelled data. So far, the players provided about 3.7M annotations and 1.7M validations, for a total of over 5.4M judgments.

Speed of annotation Over the course of the project, the games has been collecting an average of 385,000 judgments a year, i.e., slightly over 1,000 judgments per day, every day. While this is an impressive number of judgments, it only translates in an average of around 10,000 new completely annotated markables per year, or 20 new completely annotated documents, for an average of 30,000 extra words. (Progress was faster in the early years of the project, when all short documents were annotated; but as discussed in the next Section, the corpus also contains a number of fairly long texts – these are the ones still being annotated.) The project discussed in this paper was motivated by the simple calculation that at this speed, it would take us 40 years to completely annotate all the documents already in the corpus, and 300 years to completely annotated a corpus of 10M words.

		Docs	Tokens	Markables
TRAIN COMPLETE	Gutenberg	154	181142	48329 (29527)
	Wikipedia	359	244770	65050 (21803)
	Other	2	7294	2126 (1347)
	Subtotal	515	433206	115505 (52677)
TRAIN FULL	Gutenberg	194	372001	102354 (57387)
	Wikipedia	544	931752	258560 (92465)
	Other	2	7294	2128 (1347)
	Subtotal	740	1311047	363042 (151199)
DEV	Gutenberg	5	7536	2133 (1494)
	Wikipedia	35	15287	4423 (1669)
	Other	5	989	331 (126)
	Subtotal	45	23812	6887 (3289)
TEST	Gutenberg	7	20646	5925 (3332)
	Wikipedia	13	22998	7704 (3876)
	Subtotal	20	43644	13629 (7208)
All	Gutenberg	206	400183	110412 (62213)
	Wikipedia	592	970037	270687 (98010)
	Other	7	8283	2459 (1473)
	Total	805	1378503	383558 (161696)

Table 2: Summary of the current release. In parentheses the number of markables that are non-singletons.

4 Characteristics of the corpus

The *Phrase Detectives* 3.0 corpus includes all the 805 documents originally uploaded in the game. In this Section we highlight the main characteristics of the texts in this release, summarized in Table 2. For comparison, we include in the Appendix a short description of the previous release of the *Phrase Detectives* corpus, *Phrase Detectives* 2, released in 2019 (Poesio et al., 2019).

The new release The new release of the corpus, *Phrase Detectives* 3.0, is more than three times larger than the previous release of the *Phrase Detectives* corpus described in the previous section in

terms of the number of tokens (1.4M) and markables (383K). (See 'All' row in Table 2.) This makes the *Phrase Detectives* 3.0 corpus comparable in the number of tokens to ONTONOTES, but double the size of ONTONOTES in terms of markables, partly due to the singletons and non-referring expressions being included. 72% of the documents were completely annotated by the players (580 out of 805 documents), and the near totality of mentions have at least one crowd annotation (99.4%).

Genres The corpus covers mainly two genres. The Gutenberg domain consists of fiction texts from the Gutenberg Project: in part children fiction (e.g., *Alice in Wonderland*, Grimm brothers stories), in part classics (e.g., Sherlock Holmes stories). At 400K tokens, it is twice the size of the largest existing fiction corpus (Bamman et al., 2020). The Wikipedia domain consists of primarily the 'Wikipedia Unusual' documents. This subset is 1M tokens in size, substantially larger than WikiCoref (60K tokens) (Ghaddar and Langlais, 2016).

Organization The corpus is split into train, development, and test subsets, where the development and test sets are annotated by human experts (see below) and the training set is aggregated using the MPA anaphoric annotation model (Paun et al., 2018b) as described in Section 5. But crucially, two versions of the training set exists.

TRAIN COMPLETE is like the training sets released in previous versions of the corpus, in that it consists of documents that were completely annotated by the players: i.e., all markables in the documents have more than 8 judgments, and all interpretations have more than 4 validations.

The second training set, TRAIN FULL, additionally includes documents that have not yet been 'completely' annotated by the players. These documents are considerably longer, and as a result it is harder to have them completely annotated. So a state-of-the-art coreference model for this annotation scheme (Yu et al., 2020) was used in the resolve-and-aggregate setting discussed in Section 5 to augment the existing annotations by ensuring that every markable had at least one label, which would then be aggregated with the others. TRAIN FULL is three times larger than TRAIN COMPLETE, both in the number of tokens and of markables.

A New Gold The test set from the previous release of the corpus, consisting of 45 documents, is now available as DEV. DEV was fully revised by human experts for this release to correct previous labelling mistakes, and has now been annotated according to the full ARRAU guidelines, including ambiguity annotation, bridging references, and discourse deixis. In addition, a brand new TEST set of 20 documents was also created, balanced between the two domains, double in size compared to the old test set, and also annotated according to the full ARRAU guidelines by the annotators that have been preparing the ARRAU 3 release.

Domain specific training With the new release, the corpus is now large enough to be used separately for domain-specific research. We demonstrate in Section 6 that models trained on the domain-specific portion of the training set achieve comparable results to those trained on TRAIN FULL. The results indicate that the domain-specific training data can be sufficient to be used separately for dedicated research in target domains.

Long and short documents An important characteristic of the corpus is that it was designed to contain both short documents (< 2K tokens) and long ones. 34.5% of the documents are longer than 2K tokens, and the longest document reaches 14K tokens. (In contrast, in ONTONOTES only 0.4% of the documents have more than 2K tokens.) This makes our corpus a suitable resource for research on long-distance anaphora and on long document training. To this end, we use our dataset to replicate the experiments by Beltagy et al. (2020) comparing the LONGFORMER model with the ROBERTA model. In the original paper, which used the ONTONOTES corpus, no obvious differences were found between the two models, partly due to the lack of long documents. We discuss these experiments in Section 6.5. The only other corpus that we are aware of with a large portion of long documents is the CRAFT corpus (Cohen et al., 2017), which is however focused on biomedical texts.

5 Resolve-and-Aggregate

The challenge To create a reliable corpus using crowdsourcing, multiple judgments are required to ensure a good coverage of correct answers, together with sufficient evidence to enable an accurate aggregation method (Paun et al., 2018a, 2022) to distill the correct answers from the noisy ones. The problem of collecting such large number of judgments is even more serious for long documents. Annotating all anaphoric relations in long documents is chal-

lenging, partly due to the amount of time needed to complete the task, but also because the great number of entities makes it difficult for annotators to keep track of all the coreference chains. And indeed, the short documents in our corpus were completed much faster than the longer documents: the average length of the incomplete documents is 4K tokens, whereas for the complete documents is 850 tokens. Thus in our corpus the rate at which judgments are collected from players, while substantial (over 1,000 judgments per day) is not sufficient to extract reliable labels in a reasonable amount of time, as discussed in Section 3.

Possible solutions Clearly, part of the solution is to develop more engaging games, thus able to attract more players and keep them playing for longer (von Ahn and Dabbish, 2008; Jurgens and Navigli, 2014; Madge et al., 2019; Kicikoglu et al., 2019). A second ingredient is to use active learning-like approaches to minimize the number of labels required to complete the annotation (Laws et al., 2012; Li et al., 2020; Yuan et al., 2022). A number of proposals have been made in these two directions, and we are carrying out research in these areas as well (Madge et al., 2022). In this work however we investigate an approach that to our knowledge has been much less studied: combining crowdsourcing with automatic labelling. Specifically, we propose a new resolve-and-aggregate method that iteratively makes use of a coreference resolver to enhance the collected annotations. The approach is inspired, apart from *Wordrobe* (Bos et al., 2017), by previous work on Bayesian combination of classifiers (Kim and Ghahramani, 2012) which allows for aggregating predictions from classifiers and humans together with the help of a probabilistic annotation model. Both the iterative use of the coreference resolver and the application domain of the annotation model are however novel to this paper.

The coreference resolver As a coreference resolver, we use the system by Yu et al. (2020) which, to the best of our knowledge, is the only modern coreference resolver that also predicts singletons and non-referring expressions, both of which need to be annotated in our corpus. The system is an extension of (Lee et al., 2017, 2018), replacing their mention-ranking algorithm with a cluster-ranking algorithm to build the entity clusters incrementally. The system uses BERT (Devlin et al., 2019) for pretrained contextual embeddings instead of the Elmo embeddings (Peters et al., 2018) used in (Lee et al., 2018).

Aggregation Standard aggregation methods for classification labels such as the (Dawid and Skene, 1979) model are not appropriate for coreference labels, whose class space is not fixed but depends on the document mentions. However, an aggregation model for coreference judgments is now available, the mention-pair annotation model (MPA) (Paun et al., 2018b). We used MPA to aggregate judgments by players and by the coreference resolver. MPA can capture the accuracy and bias of the players, and of the coreference resolver, respectively, and adjust the aggregated labels accordingly.

Resolve-and-aggregate resolve-and-aggregate is an iterative procedure which relies on the MPA aggregation model to label the corpus, which is in turn used to retrain the coreference resolver to get better system predictions. More specifically, in the first step of the procedure we aggregate the players' annotations from the complete documents and build an initial training set, TRAIN COMPLETE. Then, we train the coreference resolver on this set, but in a gold mention setting to mimic the players who focus only on the resolution task. Having trained the system, we then use it to get predictions for the entire dataset. The resolver can be seen as a player who played all the documents in the corpus. Next, all the players' annotations and the system's predictions are aggregated using MPA, and an initial version of the entire corpus, TRAIN FULL, is built as a result. This procedure is repeated, taking TRAIN FULL as input and creating a new version every time. With each new version, the MPA-aggregated labels get refined, leading in turn to better predictions from the coreference resolver. The procedure is repeated until the performance of the resolver plateaus. The final version of the corpus contains the MPA-aggregated labels of the players' annotations and the system's best predictions. We show in the next Section that this approach results in substantial improvements in the quality of the labels produced by the coreference resolver, which translate in more accurate labels for the items not fully annotated by the players.

6 Resolving-and-Aggregating results

Experiment Setting For our experiments, we report the CoNLL F1 scores as defined in (Pradhan et al., 2012) in both singleton included and ex-

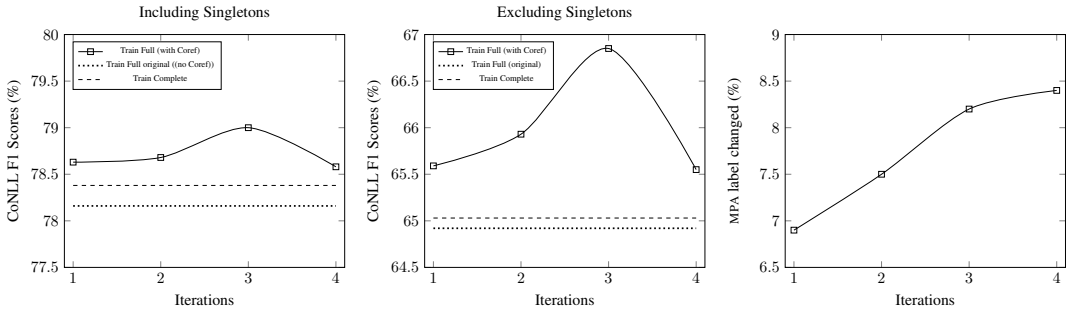

Figure 1: Left and Middle: The CoNLL scores for Yu et al. (2020) trained on different training sets and tested on the DEV set in gold mention setting. Right: The percentage of MPA labels changed by using the additional judgments from the Yu et al. (2020) system in different iterations.

cluded settings, as well as non-referring F1 scores for non-referring expressions. We use the Universal Anaphora (UA) Scorer (Yu et al., 2022b) that reports all the necessary scores.

We trained the Yu et al. (2020) system using most of its default settings. The only exception is that we always use the full context of the documents for training instead of choosing a random 1K tokens as done in Yu et al. (2020). The default setting gives priority to the short documents as for each epoch, the full context of the short documents is always used, whereas for long documents only part of the documents is used.

We establish three baselines, all using the same system Yu et al. (2020) with the same settings, but trained with different training sets. The first baseline is trained on the PREVIOUS RELEASE. The second baseline is trained on TRAIN COMPLETE (complete documents aggregated by MPA without resolve-and-aggregate). The third baseline is trained on TRAIN FULL aggregated by MPA but without annotations from the coreference resolver.

6.1 Parameter Tuning

We first trained the system using the gold mention settings to improve the quality of the corpus. We used the baseline trained on TRAIN COMPLETE to annotate the full corpus, then assigned labels to all the mentions by aggregating player and system annotations using MPA. We then trained a new model by using the full corpus (TRAIN FULL (with Coref)) and doing resolve-and-aggregate between the system and MPA in iterations until the system performance stopped improving.

The first key result is that, the system trained with TRAIN FULL (with Coref) always outperforms the baseline trained on the TRAIN COMPLETE (see Figure 1). The improvements on the singletons

Train data	CoNLL Avg. F1 Sing. (inc)	Sing. (exc)	NR F1
PREVIOUS RELEASE	65.5	53.6	36.8
TRAIN COMPLETE	66.1	54.7	39.4
TRAIN FULL(original)	64.9	52.9	35.5
TRAIN FULL(with Coref)	**66.8**	**56.1**	**40.1**
Joshi et al. (2020)	-	60.2	-

Table 3: The CoNLL and non-referring scores for (Yu et al., 2020) trained on different training sets and tested on the TEST set in predicted mention setting.

excluded setting are larger than those in the singletons included setting; this makes sense as all the models use gold mentions, hence the performance with singletons is inflated by the gold mentions. The system achieved the best performance on the third iteration with CoNLL F1 scores of 79% and 66.9% for singletons included and excluded settings respectively. This is 0.6% and 1.9% higher than the TRAIN COMPLETE baseline.

What is especially interesting is that the improvement is not just a matter of TRAIN FULL being larger than TRAIN COMPLETE: running the coreference resolver helps substantially. The system trained on TRAIN FULL original (i.e., without any automatic labels) is slightly worse than the TRAIN COMPLETE baseline, despite using the additional training data. One explanation would be that MPA's performance is affected by the lower number of judgments collected in the incomplete documents: the correct answer might not appear in the players annotations, or the players producing the annotations might not considered sufficiently reliable.

To quantify the contribution of the automatic coreference resolver, we calculate the percentage of MPA labels flipped due to the additional system annotations. We compare the labels of the TRAIN FULL (with Coref) in different iterations with the

TRAIN FULL (original) labels. We find that in the first iteration, 7% of the MPA labels (26K) were changed (see Figure 1). The percentages increased sharply until iteration 3 to 8.2% (31K) but slowed down for iteration 4. This might explain why performance starts dropping in the 4th iteration.

MPA works very well when the number of judgments is high, but performance might be affected when there are not enough annotations, e.g. for the incomplete documents. We suspected MPA might benefit more from system annotations when the document is incomplete. To assess our hypothesis, we took a closer look at MPA labels from our best iteration. We split the documents into two classes, complete and incomplete, according to our complete criterion (i.e., a document is considered complete when every markable has been annotated by at least 8 players, and each distinct interpretation has been validated by at least 4 players) and calculate a separate score for each class. We find that for the complete document only 3.3% of the MPA labels are changed as a result of the additional system annotations; in contrast, 10.8% of MPA labels are changed in the incomplete documents.

To assess the quality of these label changes, we checked the different MPA labels between iteration 3 and the original on the DEV set. Since all documents from the DEV set are complete documents, out of 7K mentions, only 201 have a different label. The TRAIN FULL (original) gets 70 of the labels correct with an accuracy of 34.8%, whereas after the 3rd iteration of resolve-and-aggregate, the number increased to 125 (62.2% accuracy). Although the sample is not large, it still gives a clear picture that even for complete documents the system annotations can improve the quality of the corpus.

6.2 Evaluation on the Test set

After finding the best setting as discussed in the previous Section, we evaluated the impact of resolve-and-aggregate on the TEST set in the more realistic predicted mention setting. As shown in Table 3, our best model trained on the the TRAIN FULL aggregated by MPA with additional coreference annotation by the Yu et al. (2020) system (TRAIN FULL(with Coref)) outperforms all the baselines in both singletons included and excluded settings. Of the baselines trained on the complete documents only, the TRAIN COMPLETE baseline works better than the PREVIOUS RELEASE baseline, most likely because the training set is larger while the quality of the annotation remains the same. But again, when training with the additional incomplete documents (TRAIN FULL (original without Coref)), the performance dropped substantially by 1%-2% when compared with the TRAIN COMPLETE baseline. This highlights again the importance of combining automatic and crowd annotations via resolve-and-aggregate: the model trained on this corpus significantly outperforms the TRAIN FULL (original) baseline by up to 3.2%. The story is the same for the models' performance on non-referring expressions (Table 3): again, the model trained on TRAIN FULL (with Coref) is top of the list.

Finally, we report the result by the Joshi et al. (2020) system on our corpus to give insight into the complexity of our corpus when compared with ONTONOTES. The system was trained on the same TRAIN FULL (with Coref) corpus. Since the Joshi et al. (2020) system only output the non-singleton clusters, we report only the CoNLL F1 score in a singleton excluded setting. As expected the system has a better CoNLL F1 score when compared with our baselines, since SpanBERT has been shown to be more effective than BERT on coreference. The Joshi et al. (2020) result on our corpus is, however, 20% lower than on ONTONOTES (79.6%), which indicates that our corpus is more complex than ONTONOTES. We hypothesize this is partially due to the longer documents and more diverse domains included in our release.

6.3 Annotation speed-up

The results showed in the previous Sections show that using automatic annotations turns the incomplete documents into documents whose quality is enough to result in improved performance when training a coreference resolver, speeding up annotation. In this section, we try to estimate the amount of time potentially saved by the proposed method. For the complete documents, we have on average 20 judgements (annotations and validations) per markable, which seems sufficient to ensure the quality of the corpus, if not perhaps necessary. For incomplete documents, the average number of judgements is currently 7.7. If we do need 20 judgements to achieve the same quality as the complete documents, we still need to collect on average 12.3 more judgements for every markable. Multiplied by the number of markables in the incomplete documents (250K), this means we would need 3M more judgements to complete all docu-

Train data	CoNLL Avg. F1 Sing. (inc)	Sing. (exc)	NR F1
Gutenberg			
DOMAIN ONLY	70.4	61.8	43.8
TRAIN FULL (with Coref)	**71.5**	**62.1**	**44.9**
Wikipedia			
DOMAIN ONLY	61.9	**50.9**	**36.1**
TRAIN FULL (with Coref)	**62.3**	50.6	36.0

Table 4: The CoNLL and non-referring scores for the system trained on different training sets and tested on the TEST set of different domains using predicted mention.

Model	Short Doc	Long Doc	All Doc
LONGFORMER	**61.0**	**67.2**	**64.7**
ROBERTA	60.1	65.2	63.1

Table 5: The CoNLL scores (exclude singletons) for LONGFORMER and ROBERTA trained on TRAIN FULL and tested on the TEST set using gold mentions.

ments in the game. In the last five years, we have been averaging 334K judgements per year, which means if we proceed at the current speed, we need another 9 years before we can release this corpus. In other words, the resolve-and-aggregate method significantly speeds up the annotation process.

6.4 Domain-specific Training

Thanks to the resolve-and-aggregate method, this new release gives us datasets of a reasonable size for both the Gutenberg (fiction) and Wikipedia domains. We evaluated system performance on the domain-specific portion - e.g., for Fiction we trained our model on the Gutenberg section of TRAIN FULL (with Coref) and tested it on the Gutenberg section of the TEST. We then compared the performance of these domain-specific models with that of the best system trained on the entire corpus. As shown in Table 4, the DOMAIN ONLY systems trained on the domain-specific subsections of the corpus achieve scores close to the system trained on the full corpus. This suggests each domain-specific part of the corpus is sufficiently large to be used for domain-specific research.

6.5 Long and short documents

As stated earlier, one of the emerging challenges for research on anaphora (and NLP in general) are longer documents (>2K tokens). Our corpus is unusual in that it includes a large number of documents more than 2K in length, with the longest document containing 14K tokens. TEST also balances short (55%) and long (45%) documents.

To test that the corpus can support research on anaphora in long documents, we used it to replicate the comparison in (Beltagy et al., 2020) between their new model designed specifically for longer documents, the LONGFORMER, with ROBERTA (Liu et al., 2019). In that paper, the LONGFORMER is compared with ROBERTA on the ONTONOTES corpus, without however finding a clear difference between the two systems. We suspected this might be because ONTONOTES does not contain enough long documents to observe improvements. We replicated the experiments by Beltagy et al. with our corpus, and report the CoNLL F1 score on full TEST as well as separate scores for long/short documents. (Since neither system predicts singletons and non-referring expressions, we report the CoNLL F1 scores in the singleton excluded setting.) We evaluated the systems with the gold mentions so that the system's performance will not be affected by mention detection.

Table 5 shows the results for both systems on different test set. The LONGFORMER works better on all test sets, but with a much larger gain over ROBERTA on long documents: the improvement over ROBERTA is 0.9% and 2% on short and long documents respectively. This finding confirms that long documents benefit more from the LONGFORMER architecture, while also showing that our corpus can be used to differentiate systems designed to perform on long documents.

7 Conclusions

This research makes two main contributions. First of all, we proposed an iterative method for speeding up anaphoric annotation via GWAPs by combining crowdsourced data with labels produced by an automatic coreference resolver, and aggregating the labels using a probabilistic annotation method; and showed that the resulting extension leads to quantifiable improvements in model performance. The method can be easily extended to other types of annotation. Second, we introduced a new corpus for anaphoric reference which, thanks to the use of resolve-and-aggregate, is of a comparable size to ONTONOTES in terms of tokens, but twice the size in terms of markables; it contains two substantial datasets for genres not covered in ONTONOTES; and it includes both short and long documents. The corpus will be made freely available with all judgements.

8 Limitations

The main limitation of this work is that the new release is still only twice the size of ONTONOTES in terms of markables. In ongoing work, we are developing a new platform to label a corpus twenty times the size of the current release. The new platform combines more engaging games with active-learning like methods for allocating work to players more efficiently and according to their linguistic understanding. We hope that the new platform, in combination with the methods proposed here, will allow us to label the new and larger dataset much more quickly.

A second limitation of the new release is that the markables in the corpus were automatically extracted; thus, the quality of the mentions is lower than in corpora in which they were hand-identified. The approach followed in these years has been to ask our players to signal issues; as a result, tens of thousands of markables were hand-corrected. However, this approach doesn't really lend itself to scaling up. Thus, in our new platforms we are following a different strategy: asking our players to do the corrections themselves, by including also games to check other levels of linguistic interpretation.

A third limitation, in particular in comparison with ONTONOTES, is that this release of the corpus only contains English documents, although a small amount of Italian documents was uploaded in the game.

Acknowledgements

This research was supported in part by the ANAWIKI project, funded by EPSRC (EP/F00575X/1);[4] in part by the DALI project, funded by the European Research Council (ERC), Grant agreement ID: 695662;[5] in part by the ARCIDUCA project, funded by EPSRC (EP/W001632/1).

References

Abdulrahman Aloraini and Massimo Poesio. 2021. Data augmentation methods for anaphoric zero pronouns. In *Proc. of the CRAC Workshop*.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proc. of LREC*. European Language Resources Association (ELRA), Association for Computational Linguistics (ACL).

Valerio Basile, Johan Bos, Kilian Evang, and Noortje Venhuizen. 2012. Developing a large semantically annotated corpus. In *Proc. of LREC*, pages 3196–3200, Istanbul, Turkey.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Federico Bonetti and Sara Tonelli. 2021. Challenges in designing games with a purpose for abusive language annotation. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*.

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. 2016. *Harnessing Collective Intelligence on Social Networks*. Ph.D. thesis, University of Essex, School of Computer Science and Electronic Engineering.

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

Jon Chamberlain, Massimo Poesio, and Udo Kruschwitz. 2016. Phrase detectives corpus 1.0: Crowdsourced anaphoric coreference. In *Proceedings of LREC*, Portoroz, Slovenia.

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.

Christopher Cieri, James Fiumara, and Jonathan Wright. 2021. Using games to augment corpora for language recognition and confusability. In *Proc. of Interspeech: 22nd Annual Conference of the International Speech Communication*.

K Bretonnel Cohen, Arrick Lanfranchi, Miji Joo-young Choi, Michael Bada, William A Baumgartner, Natalya Panteleyeva, Karin Verspoor, Martha Palmer, and Lawrence E Hunter. 2017. Coreference annotation and resolution in the colorado richly annotated full text (craft) corpus of biomedical journal articles. *BMC bioinformatics*, 18(1):1–14.

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.

[4] https://anawiki.essex.ac.uk/

[5] http://www.dali-ambiguity.org

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Anca Dumitrache, Lora Aroyo, Chris Welty, Robert-Jan Sips, and Anthony Levas. 2013. Dr. detective: combining gamification techniques and crowdsourcing to create a gold standard in medical text. In *Proc. of CrowdSem*.

Dagmara Dziedzic and Wojciech Włodarczyk. 2017. Making nlp games with a purpose fun to play using free to play mechanics: RoboCorp case study. In *Proc. of the Games4NLP Symposium*.

Ali Emami, Paul Trichelair, Adam Trischler, Kaheer Suleman, Hannes Schulz, and Jackie Chi Kit Cheung. 2019. The knowref coreference corpus: Removing gender and number cues for difficult pronominal anaphora resolution. In *Proc. of the ACL*.

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.

Luke Gessler, Siyao Peng, Yang Liu, Yilun Zhu, Shabnam Behzad, and Amir Zeldes. 2020. AMALGUM – a free, balanced, multilayer english web corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, page 5267–5275. European Language Resources Association.

Abbas Ghaddar and Phillippe Langlais. 2016. WikiCoref: An English coreference-annotated corpus of Wikipedia articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 136–142, Portorož, Slovenia. European Language Resources Association (ELRA).

Iris Hendrickx, Gosse Bouma, Frederik Coppens, Walter Daelemans, Veronique Hoste, Geert Kloosterman, Anne-Marie Mineur, Joeri Van Der Vloet, and Jean-Luc Verschelde. 2008. A coreference corpus and resolution system for dutch. In *Proc. LREC*.

Erhard W. Hinrichs, Sandra Kübler, 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.

Barbora Hladká, Jiří Mírovský, and Pavel Schlesinger. 2009. Designing a language game for collecting coreference annotation. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 52–55, Singapore.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Alain Joubert, Mathieu Lafourcade, and Nathalie Le Brun. 2018. The jeuxdemots project is 10 years old: What we have learned. In *Proceedings of the 2018 LREC Workshop on Games and Gamification for Natural Language Processing (Games4NLP)*, pages 22–26.

David Jurgens and Roberto Navigli. 2014. It's all fun and games until someone annotates: Video games with a purpose for linguistic annotation. *Transactions of the ACL*.

Doruk Kicikoglu, Richard Bartle, Jon Chamberlain, and Massimo Poesio. 2019. Wormingo: a'true gamification' approach to anaphoric annotation. In *FDG'19: Proc. of the GAMNLP Workshop at the 14th International Conference on the Foundations of Digital Games*, pages 1–7, San Luis Obispo.

Hyun-Chul Kim and Zoubin Ghahramani. 2012. Bayesian classifier combination. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 619–627, La Palma, Canary Islands. PMLR.

Markus Krause, Aneta Takhtamysheva, Marion Wittstock, and Rainer Malaka. 2010. Frontiers of a paradigm: exploring human computation with digital games. In *Proceedings of HCOMP - the ACM SIGKDD Workshop on Human Computation*, pages 22–25.

Mathieu Lafourcade. 2007. Making people play for lexical acquisition with the JeuxDeMots prototype. In *Proc. of SNLP*.

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

Fred Landragin. 2016. Description, modéelisation et détection automatique des chaînes de réeférence (democrat). *Bulletin de l'Association Française pour l'Intelligence Artificielle*, 92:11–15.

Florian Laws, Florian Heimerl, and Hinrich Schütze. 2012. Active learning for coreference resolution. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 508–512. Association for Computational Linguistics.

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*.

Belinda Z. Li, Gabriel Stanovsky, and Luke Zettlemoyer. 2020. Active learning for coreference resolution using discrete annotation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 8320–8331. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Chris Madge, Richard Bartle, Jon Chamberlain, Udo Kruschwitz, and Massimo Poesio. 2019. Incremental game mechanics applied to text annotation. In *CHI PLAY 2019 - Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, pages 545–558, Barcelona.

Chris Madge, Jussi Brightmore, Doruk Kicikoglu, Fatima Althani, Richard Bartle, Jon Chamberlain, Udo Kruschwitz, and Massimo Poesio. 2022. Lingotowns: A virtual world for natural language annotation and language learning. In *CHI PLAY 2022 - Proceedings of the Annual Symposium on Computer-Human Interaction in Play, Demo session*, Bremen.

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.

Judith Muzerelle, Anaïs Lefeuvre, Emmanuel Schang, Jean-Yves Antoine, Aurore Pelletier, Denis Maurel, Iris Eshkol, and Jeanne Villaneau. 2014. Ancor_centre, a large free spoken french coreference corpus. In *Proc. of LREC*.

Anna Nedoluzhko, Michal Novák, Silvie Cinková, Marie Mikulová, and Jiří Mírovský. 2016. Coreference in prague czech-english dependency treebank. In *Proc. of LREC*, Portoroz. ELRA.

Silviu Paun, Ron Artstein, and Massimo Poesio. 2022. Statistical methods for annotation analysis. *Synthesis Lectures on Human Language Technologies*, 15(1):1–217.

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*.

Massimo Poesio, Jon Chamberlain, and Udo Kruschwitz. 2017. Crowdsourcing. In N. Ide and J. Pustejovsky, editors, *The Handbook of Annotation*, pages 277–295. Springer.

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Silviu Paun, Alexandra Uma, and Juntao Yu. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proc. of NAACL*, page 1778–1789, Minneapolis. Association for Computational Linguistics (ACL).

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, 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.

Ellen F. Prince. 1992. The ZPG letter: subjects, definiteness, and information status. In S. Thompson and W. Mann, editors, *Discourse description: diverse analyses of a fund-raising text*, pages 295–325. John Benjamins.

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

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. 2020. Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU corpus. *Journal of Natural Language Engineering*.

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

Luis von Ahn and Laura Dabbish. 2008. General techniques for designing games with a purpose. *Communications of the ACM*, pages 58–67.

Sam J. Wiseman, Alexander M. Rush, Stuart M. Shieber, and Jason Weston. 2015. Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proc. of the ACL*, Bejing.

Patrick Xia and Benjamin Van Durme. 2021. Moving on from ontonotes: Coreference resolution model transfer. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, page 5241–5256. Association for Computational Linguistics.

Juntao Yu, Sopan Khosla, Ramesh Manuvinakurike, Lori Levin, Vincent Ng, Massimo Poesio, Michael Strube, and Massimo Poesio. 2022a. The CODI/CRAC 2022 shared task on anaphora resolution, bridging and discourse deixis in dialogue. In *Proc. of CODI/CRAC Shared Task*.

Juntao Yu, Sopan Khosla, Nafise Sadat Moosavi, Silviu Paun, Sameer Pradhan, and Massimo Poesio. 2022b. The universal anaphora scorer. In *Proceedings of the Language Resources and Evaluation Conference*, pages 4873–4883, Marseille, France. European Language Resources Association.

Juntao Yu, Alexandra Uma, and Massimo Poesio. 2020. A cluster ranking model for full anaphora resolution. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 11–20, Marseille, France. European Language Resources Association.

Michelle Yuan, Patrick Xia, Chandler May, Benjamin Van Durme, and Jordan Boyd-Graber. 2022. Adapting coreference resolution models through active learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, page 7533–7549. Association for Computational Linguistics.

Amir Zeldes. 2020. *Multilayer corpus studies*. Routledge.

Appendix

A The previous release of the corpus

Phrase Detectives 2 consisted of a total of 542 documents containing 408K tokens and 108K markables from two main genres: Wikipedia articles and fiction from the Gutenberg collection. This version of the corpus was divided in two subsets. The subset referred to to as PD_{silver} consisted 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 is also provided. 45 additional documents were also gold-annotated by two experts annotators. The subset of the corpus for which both gold and silver annotations are available was called PD_{gold}, as it is intended to be used as test set.[6] The gold subset consists of a total of 23K tokens and 6K markables. The contents of the *Phrase Detectives* 2 corpus are summarized in Table 6.

		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)
PD_{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 6: Summary of the contents of the 2019 release of the *Phrase Detectives* corpus. The numbers in parentheses indicate the total number of markables that are non-singletons.

B Detailed Evaluation Results

This appendix section includes the detailed evaluation results for this paper. More specifically, Table 7 and Table 8 show the detailed scores for our experiments on predicted mentions (discussed in Section 6.2); Table 9 and Table 10 show the detailed scores of coreference and non-referring expressions for the domain specific training experiments set out in Section 6.4. Table 11 shows the detailed scores for the long/short documents experiments discussed in Section 6.5.

Singletons	Train Data	MUC			BCUB			CEAFE			Avg. F1
		P	R	F1	P	R	F1	P	R	F1	
Included	PREVIOUS RELEASE	83.2	60.6	70.1	73.9	54.8	62.9	59.7	67.5	63.4	65.5
	TRAIN COMPLETE	83.1	62.1	71.1	74.7	54.5	63.0	62.4	66.2	64.3	66.1
	TRAIN FULL (original)	84.2	58.6	69.1	76.0	52.5	62.1	60.2	66.8	63.4	64.9
	TRAIN FULL (with Coref)	83.4	63.4	72.0	74.5	55.3	63.5	63.4	66.5	64.9	66.8
Excluded	PREVIOUS RELEASE	83.2	60.6	70.1	72.4	36.6	48.6	52.8	34.9	42.0	53.6
	TRAIN COMPLETE	83.1	62.1	71.1	71.6	37.6	49.3	53.6	36.8	43.6	54.7
	TRAIN FULL (original)	84.2	58.6	69.1	73.4	33.4	46.0	55.1	36.1	43.7	52.9
	TRAIN FULL (with Coref)	83.4	63.4	72.0	71.0	39.0	50.4	56.2	38.8	45.9	56.1
	SpanBERT-Large (Joshi et al.)	89.0	65.5	75.5	79.7	43.2	56.0	60.9	41.4	49.2	60.2
	SpanBERT-Base (Joshi et al.)	88.1	64.6	74.5	79.3	43.4	56.1	58.8	40.2	47.7	59.4

Table 7: The CoNLL scores for the Yu et al. (2020) and Joshi et al. (2020) systems trained on different training sets and tested on the TEST set in predicted mention setting.

[6] PD_{gold} is the dataset released in 2016 as *Phrase Detectives* corpus, Release 1 (Chamberlain et al., 2016).

Train data	P	R	F1
PREVIOUS RELEASE	73.8	24.6	36.8
TRAIN COMPLETE	71.1	27.3	39.4
TRAIN FULL(original)	75.1	23.3	35.5
TRAIN FULL(with Coref)	77.9	27.0	40.1

Table 8: Non-referring scores for Yu et al. (2020) system trained on different training sets and tested on the TEST set in predicted mention setting.

Singletons	Train Data	MUC			BCUB			CEAFE			Avg. F1
		P	R	F1	P	R	F1	P	R	F1	
					Gutenberg						
Included	DOMAIN ONLY	87.3	75.2	80.8	71.5	57.6	63.8	61.3	73.4	66.8	70.4
	TRAIN FULL (with Coref)	87.6	75.9	81.3	73.2	57.8	64.6	65.0	72.5	68.5	71.5
Excluded	DOMAIN ONLY	87.3	75.2	80.8	70.2	42.0	52.5	56.3	48.5	52.1	61.8
	TRAIN FULL (with Coref)	87.6	75.9	81.3	69.9	42.7	53.0	56.3	48.2	51.9	62.1
					Wikipedia						
Included	DOMAIN ONLY	75.4	52.0	61.6	72.8	54.0	62.0	61.2	62.8	62.0	61.9
	TRAIN FULL (with Coref)	78.1	51.3	61.9	75.5	53.3	62.5	62.4	62.7	62.5	62.3
Excluded	DOMAIN ONLY	75.4	52.0	61.6	69.8	36.9	48.3	54.1	35.5	42.9	50.9
	TRAIN FULL (with Coref)	78.1	51.3	61.9	72.3	35.8	47.9	56.2	33.5	42.0	50.6

Table 9: The CoNLL scores for Yu et al. (2020) system trained on different training sets and tested on the TEST set of different domains in predicted mention setting.

Train data	P	R	F1
	Gutenberg		
DOMAIN ONLY	79.9	30.2	43.8
TRAIN FULL (with Coref)	84.0	30.6	44.9
	Wikipedia		
DOMAIN ONLY	71.6	24.1	36.1
TRAIN FULL (with Coref)	72.4	24.0	36.0

Table 10: Non-referring scores for Yu et al. (2020) system trained on different training sets and tested on the TEST set of different domains in predicted mention setting.

Settings	Model	MUC			BCUB			CEAFE			Avg. F1
		P	R	F1	P	R	F1	P	R	F1	
Short Doc	LONGFORMER	96.2	61.5	75.0	88.4	45.9	60.4	74.8	34.7	47.4	61.0
	ROBERTA	96.3	60.8	74.5	89.3	45.1	59.9	71.1	33.9	45.9	60.1
Long Doc	LONGFORMER	94.2	71.6	81.3	77.6	53.3	63.1	73.2	46.7	57.0	67.2
	ROBERTA	94.4	71.1	81.1	76.3	49.4	59.9	71.9	43.8	54.4	65.2
All Doc	LONGFORMER	94.9	67.5	78.9	81.6	50.2	62.2	73.8	41.3	52.9	64.7
	ROBERTA	95.1	66.9	78.5	81.1	47.6	60.0	71.6	39.3	50.7	63.1

Table 11: The CoNLL scores for LONGFORMER and ROBERTA systems trained on TRAIN FULL and tested on the TEST set using gold mentions in a singleton excluded setting.