The CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue

Sopan Khosla[1,7]*, **Juntao Yu**[2], **Ramesh Manuvinakurike**[3], **Vincent Ng**[4], **Massimo Poesio**[5], **Michael Strube**[6], and **Carolyn Rosé**[1]
[1]Canegie Mellon Univ., USA; [2]Univ. of Essex, UK; [3]Intel Labs, USA; [4]UT Dallas, USA; [5]Queen Mary Univ., UK; [6]HITS, Germany; [7]AWS AI, Amazon, USA
sopankh@amazon.com; j.yu@essex.ac.uk; ramesh.manuvinakurike@intel.com; vince@hlt.utdallas.edu; m.poesio@qmul.ac.uk; Michael.Strube@h-its.org; cprose@cs.cmu.edu

Abstract

In this paper, we provide an overview of the CODI-CRAC 2021 Shared Task. The shared task focuses on detecting anaphoric relations in different genres of conversations. Using five conversational datasets, four of which have been newly annotated with a wide range of anaphoric relations: identity, bridging references and discourse deixis, we defined multiple tasks focusing individually on these key relations. We discuss the evaluation scripts used to assess the system performance on these tasks, and provide a brief summary of the participating systems and the results obtained across 115 runs from six teams, with most submissions achieving significantly better results than our baseline methods.

1 Introduction

The performance of models for single-antecedent anaphora resolution on the aspects of anaphoric interpretation annotated in the standard ONTONOTES dataset (Pradhan et al., 2012) has greatly improved in recent years (Wiseman et al., 2015; Lee et al., 2017, 2018; Kantor and Globerson, 2019; Joshi et al., 2020). So the attention of the community has started to turn to more complex cases of anaphora not found or not properly tested in ONTONOTES.

Well-known examples of this trend are work on the cases of anaphora whose interpretation requires some form of commonsense knowledge tested by benchmarks for the Winograd Schema Challenge (Rahman and Ng, 2012; Liu et al., 2017; Sakaguchi et al., 2020), or the pronominal anaphors that cannot be resolved purely using gender, for which benchmarks such as GAP have been developed (Webster et al., 2018). GAP, however, still focused on identity coreference.

In addition, more research has been carried out on aspects of anaphoric interpretation that go beyond identity anaphora but are covered by datasets such as ARRAU (Poesio et al., 2018; Uryupina et al., 2020). These include, e.g., bridging reference (Clark, 1977; Hou et al., 2018; Hou, 2020; Yu and Poesio, 2020; Kobayashi and Ng, 2021), discourse deixis (Webber, 1991; Marasović et al., 2017; Kolhatkar et al., 2018) or split-antecedent anaphora (Eschenbach et al., 1989; Vala et al., 2016; Zhou and Choi, 2018; Yu et al., 2020b, 2021).

There has been interest in other genres apart from news. This includes substantial research on annotating and resolving coreference in biomedical and other scientific domains (Cohen et al., 2017; Lu and Poesio, 2021) as well as in literary documents (Bamman et al., 2020). There are, however, language genres still understudied in the literature on anaphoric reference. Arguably the most important among these is conversational language in dialogue. Anaphora resolution in dialogue requires systems to handle grammatically incorrect language suffering from disfluencies. Dialogue involves much more deictic reference, vaguer anaphoric and discourse deictic reference, speaker grounding of pronouns and long-distance conversation structure. These are complexities that are often missing in news or Wikipedia articles, which form a large chunk of current datasets for coreference resolution. There has been some research on coreference in dialogue (Byron, 2002; Eckert and Strube, 2001; Müller, 2008), but very limited in scope (primarily related to pronominal interpretation), due to the lack of suitable corpora. The one language for which substantial corpora of coreference in dialogue exist is French: the ANCOR corpus (Muzerelle et al., 2014) has enabled the development of an end-to-end neural model for coreference interpretation in dialogue by Grobol (2020). For English, the one resource we are aware of fully annotated for anaphoric reference is the TRAINS corpora included in the ARRAU corpus (Uryupina et al., 2020).

The objective of the CODI-CRAC 2021 Shared

*Work done when the author was a student at CMU

Task in Anaphora Resolution in Dialogue[1] was to provide participants with the opportunity to develop automated approaches for coreference resolution that tackle less studied forms of anaphora and generalize on different types of conversational setups. Specifically, the shared task is divided in three tasks that individually tackle a particular anaphoric relation: identity, bridging, and discourse deixis. To evaluate the participating systems, we provide development and test sets consisting of four conversational datasets from different domains newly annotated with the above-mentioned relations. This lack of in-domain training data also poses research challenges related to transfer learning and out-of-domain generalization of learned representations. To accommodate for systems that use gold/predicted mentions for bridging and discourse deixis tasks, we set up separate leaderboards for the two settings.

Our goal in this paper is to present an overview of the CODI-CRAC 2021 shared task. We begin by providing some background in Section 2 and introducing the new CODI-CRAC 2021 corpus in Section 3. We then provide an extensive overview of the different CODI-CRAC 2021 tasks, markable settings, and evaluation metrics in Section 4, and submission details in Section 5. This is followed by details of the baselines in Section 6 and participating systems in Section 7. We present a discussion of the performance of the systems on different tasks and sub-corpora in Section 8, and finally conclude this paper in Section 9.

2 Background

2.1 Beyond Identity Coreference

Most modern anaphoric annotation projects cover basic identity anaphora as in (1).

(1) [Mary]$_i$ bought [a new dress]$_j$ but [it]$_j$ didn't fit [her]$_i$.

However, many other types of identity anaphora exist, as well as other types of anaphoric relations that are not annotated in ONTONOTES but are annotated in other corpora.

Split-antecedent anaphora In ONTONOTES, plural reference is only marked when the antecedent is mentioned by a single noun phrase. However, **split-antecedent anaphors** are also possible (Eschenbach et al., 1989; Kamp and Reyle, 1993), as in (2). These are also cases of plural identity coreference, but to sets composed of two or more entities introduced by separate noun phrases. Such references are annotated in, e.g., ARRAU (Uryupina et al., 2020), GUM (Zeldes, 2017) and *Phrase Detectives* (Poesio et al., 2019).

(2) [John]$_1$ met [Mary]$_2$. [He]$_1$ greeted [her]$_2$. [They]$_{1,2}$ went to the movies.

Discourse deixis In ONTONOTES, **event anaphora**, a subtype of **discourse deixis** (Webber, 1991; Kolhatkar et al., 2018) is marked, as in (3) (where *[that]* arguably refers to the event of a white rabbit with pink ears running past Alice) but not the whole range of abstract anaphora, illustrated by, e.g., (4), where again arguably *[this]* refers to the fact that the Rabbit was able to talk. (Both examples from the *Phrase Detectives* corpus (Poesio et al., 2019).)

(3) So she was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly a White Rabbit with pink eyes ran close by her. There was nothing so VERY remarkable in [that]; nor did Alice think it so VERY much out of the way to hear the Rabbit say to itself, 'Oh dear! Oh dear! I shall be late!' (when she thought it over afterwards, it occurred to her that she ought to have wondered at this, but at the time it all seemed quite natural); ...

(4) There was nothing so VERY remarkable in that; nor did Alice think it so VERY much out of the way to hear the Rabbit say to itself, 'Oh dear! Oh dear! I shall be late!' (when she thought it over afterwards, it occurred to her that she ought to have wondered at [this], but at the time it all seemed quite natural); but when the Rabbit actually TOOK A WATCH OUT OF ITS WAISTCOAT-POCKET, and looked at it, and then hurried on, Alice started to her feet, for it flashed across her mind that she had never before seen a rabbit with either a waistcoat-pocket, or a watch to take out of it, and burning with curiosity, she ran across the field after it, and fortunately was

[1] https://competitions.codalab.org/competitions/30312

just in time to see it pop down a large rabbit-hole under the hedge.

Bridging references There are other forms of anaphoric reference besides identity, and there are now a number of corpora annotating (a subset of) these forms. Possibly the most studied of non-identity anaphora is **bridging reference** or **associative anaphora** (Clark, 1977; Hawkins, 1978; Prince, 1981) as in (5), where bridging reference / associative anaphora *the roof* refers to an object which is related to / associated with, but not identical to, the *hall*. We also take bridging reference to cover *other* anaphora as in (6), as well as other cases of association such as identity of sense anaphora, etc. (Poesio, 2016).

(5) There was not a moment to be lost: away went Alice like the wind, and was just in time to hear it say, as it turned a corner, 'Oh my ears and whiskers, how late it's getting!' She was close behind it when she turned the corner, but the Rabbit was no longer to be seen: she found herself in [a long, low hall, which was lit up by a row of lamps hanging from [the roof]].
There were doors all round the hall, but they were all locked; and when Alice had been all the way down one side and up the other, trying every door, she walked sadly down [the middle], wondering how she was ever to get out again.

(6) There was not a moment to be lost: away went Alice like the wind, and was just in time to hear it say, as it turned a corner, 'Oh my ears and whiskers, how late it's getting!' She was close behind it when she turned the corner, but the Rabbit was no longer to be seen: she found herself in a long, low hall, which was lit up by a row of lamps hanging from the roof.
There were doors all round the hall, but they were all locked; and when Alice had been all the way down [one side] and up [the other], trying every door, she walked sadly down the middle, wondering how she was ever to get out again.

2.2 The CRAC 2018 Shared Task

The more general types of anaphoric reference just discussed are now routinely annotated in a number of corpora, including ANCORA (Recasens and Martí, 2010), ARRAU (Uryupina et al., 2020), GNOME (Poesio, 2004), GUM (Zeldes, 2017), ISNOTES (Markert et al., 2012), the Prague Dependency Treebank (Nedoluzhko, 2013), and TÜBADZ (Versley, 2008). (See Poesio et al. (2016) for a more detailed survey and Nedoluzhko et al. (2021) for a more recent, extensive update.)

Some of these resources are of a sufficient size to support shared tasks. In particular, the ARRAU corpus was used as the dataset for the Shared Task on Anaphora Resolution with ARRAU in the CRAC 2018 Workshop (Poesio et al., 2018). That shared task was articulated around three tasks: identity coreference (including identification of non-referring expressions), bridging references, and discourse deixis. The organization of the shared task resulted in the development of an extended version of the Coreference Reference Scorer (Pradhan et al., 2014), which also scores non-referring expressions. Separate scorers were developed for bridging reference resolution, carrying out both mention-based evaluation and entity-based evaluation of bridging references, as done by Hou et al. (2018), and for discourse deixis, based on Kolhatkar and Hirst (2014). The present shared task was modeled on that.

2.3 Universal Anaphora

In order to enable further progress in the empirical study of anaphora by coordinating the many existing efforts to annotate not just identity coreference, but all aspects of anaphoric interpretation from identity of sense anaphora to bridging to discourse deixis; and not just for English, but all languages, the **Universal Anaphora** (UA) initiative was launched in 2020.[2] Progress so far includes a first proposal concerning the range of phenomena to be covered, as well as a survey of the range of existing anaphoric annotations and a proposal for a markup format extending the CONLL-U format developed by the **Universal Dependencies** initiative[3] with mechanisms for marking up the range of anaphoric information covered by UA. Crucially, a scorer able to evaluate all types of anaphoric reference in the scope of the proposal was also developed, which we used in this shared task. The scorer is briefly discussed in Section 4.2.

[2] https://universalanaphora.github.io/UniversalAnaphora/

[3] https://universaldependencies.org/

2.4 Datasets of Anaphora in Dialogue

A limitation of most resources annotated for anaphora is that they mostly focus on expository text. The one substantial dataset of anaphoric relations in dialogue is ANCOR for French (Muzerelle et al., 2014), in which identity and bridging anaphora are annotated. Among the small number of English corpora that cover dialogue include ONTONOTES (Pradhan et al., 2012), which contains a small number of conversations annotated for identity anaphora and a small subtype of discourse deixis (as discussed earlier). ARRAU's (Poesio and Artstein, 2008; Uryupina et al., 2020) TRAINS sub-corpus consists of task-oriented dialogues for identity, bridging, and discourse deixis. We include TRAINS in CODI-CRAC 2021 training data. The more recently released ONTOGUM (Zhu et al., 2021) builds upon the ONTONOTES schema and adds several new genres (including more spoken data) to the ONTONOTES family. Both identity anaphora and bridging are annotated in the dataset.

3 The CODI-CRAC 2021 Corpus

One of the objectives of the CODI-CRAC 2021 Shared Task was to annotate new data for studying anaphora in dialogue. The only existing dataset covering the full range of phenomena and with some coverage of dialogue, the ARRAU data used for the CRAC 2018 Shared Task, was used as training material. In addition, new data from dialogue corpora were annotated for development and testing using the same annotation scheme used in ARRAU.

3.1 ARRAU: Corpus and Annotation Scheme

Genres The ARRAU corpus[4] (Poesio and Artstein, 2008; Uryupina et al., 2020) was designed to cover a variety of genres. It includes a substantial amount of news text in a sub-corpus called RST, consisting of the entire subset of the Penn Treebank (Marcus et al., 1993) that was annotated in the RST treebank (Carlson et al., 2003). In addition to the news data, ARRAU includes three more sub-corpora. The TRAINS sub-corpus includes all the task-oriented dialogues in the TRAINS-93 corpus[5] as well as the pilot dialogues in the so-called TRAINS-91 corpus. The PEAR sub-corpus consists of the complete collection of spoken narratives in the Pear Stories that provided some of the early evidence on salience and anaphoric reference (Chafe, 1980), and the GNOME sub-corpus covers documents from the medical and art history genres covered by the GNOME corpus (Poesio, 2004). The same coding scheme was used for all sub-corpora, but separate guidelines were written for the textual and the spoken dialogue sub-corpora. RST, TRAINS-93 and PEAR were used for the CRAC 2018 shared task. For this year's shared task, TRAINS-91 was used for development in the trial phase, whereas all other datasets were used for training.[6]

Annotation scheme The original annotation scheme used for Release 1 (Poesio and Artstein, 2008) is distributed with the dataset and is also available from the ARRAU corpus page. For the second release (Uryupina et al., 2020), the guidelines for bridging were extended and genericity was also annotated using the GNOME guidelines, but a complete new manual was not produced. However, a fairly extensive description can be found in Uryupina et al. (2020). Following the CRAC 2018 shared task, the guidelines for bridging, semantic category and genericity were further revised as part of the work on ARRAU Release 3, which is almost ready, and a full revision of the annotation manual was also started. These guidelines were used in the current annotation. A brief summary follows.

Markable definition Many, especially among the older, anaphorically annotated corpora impose syntactic, semantic or discourse-based restrictions on markables. For instance, in ONTONOTES neither expletives nor singletons are annotated (for a discussion of the state of the art in anaphoric annotation, see Poesio et al. (2016)). By contrast, in ARRAU *all* NPs are considered as markables, including non-referring expressions (e.g., expletives such as *it* or predicative NPs such as *a busy place*) in (7), and expressions do not corefer with any other markable and thus form a singleton coreference chain. Moreover, in ARRAU non-referring markables are manually sub-classified into expletives, predicative, and quantifiers. In addition, all generic references are marked, including premodifiers when the entity referred to is mentioned again, e.g., in the case of

[4] http://www.arrauproject.org

[5] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC95S25

[6] The original intention had been to use the soon-to-be-released ARRAU 3, but as the work on this version was still under way by the time the training data had to be released, ARRAU 2 was used instead–i.e., the exact same version used for the CRAC 2018 shared task.

the proper name *US* in (8), and premodifiers that refer to a kind, like *exchange-rate* in (9).

(7) [It] seems to be [a busy place]

(8) . . . The Treasury Department said that the [US]$_1$ trade deficit may worsen next year after two years of significant improvement. . . The statement was the [US]$_1$'s government first acknowledgment of what other groups, such as the International Monetary Fund, have been predicting for months.

(9) The Treasury report, which is required annually by a provision of the 1988 trade act, again took South Korea to task for its [exchange-rate]$_1$ policies. "We believe there have continued to be indications of [exchange-rate]$_1$ manipulation . . .

Types of anaphoric relations marked The ARRAU guidelines support annotation of different types of anaphoric relations. All referring markables are marked as either `discourse new` or `discourse old`. Discourse new mentions introduce new entities and thus are not marked as being coreferent with an entity already introduced (**antecedent**). For discourse-old mentions, an antecedent can be identified, either of type `phrase` (if the antecedent was introduced using a nominal markable) or `segment` (not introduced by a nominal markable, for **discourse deixis**).[7] In addition, referring NPs can be marked as **related** to a previously mentioned discourse entity to identify them as examples of associative (**bridging**) anaphora.

Bridging references Annotating — indeed, even identifying — bridging references in a reliable way is difficult, which is one of the reasons why so few large-scale corpora for anaphora include this type of annotation (Poesio et al., 2016; Kobayashi and Ng, 2020). The ARRAU guidelines for bridging anaphora are based on experiments that were started by Vieira and Poesio (Poesio and Vieira, 1998) and continued in the GNOME project (Poesio, 2004). The ARRAU Release 1 guidelines followed the GNOME guidelines, but with an extension and a simplification. Annotators were asked to mark a markable as `related` to a particular antecedent if it stood to that antecedent in one of the relations identified in GNOME (indeed, the same examples were used), and in addition, if they stood in two additional relations (but without testing the reliability of this annotation):

- `other`, for *other* NPs, broadly following the guidelines in Modjeska (2003);
- an `undersp-rel` relation for 'obvious cases of bridging that didn't fit any other category'.

Discourse deixis Discourse deixis in its full form is a very complex form of reference, both to annotate (Kolhatkar et al., 2018) and to resolve. Very few anaphoric annotation projects have attempted to annotate discourse deixis in its entirety (Kolhatkar et al., 2018). More typical is a partial annotation, as in (Byron and Allen, 1998; Navaretta, 2000), who annotated pronominal reference to abstract objects; in ONTONOTES, where event anaphora was marked (Pradhan et al., 2007); and in the work of Kolhatkar and Hirst (2014), which focused on so-called shell nouns. In ARRAU, A coder specifying that a referring expression is discourse-old is asked whether its antecedent was introduced using a `phrase` (markable) or a `segment` (discourse segment). Coders who choose `segment` have to mark a sequence of *predefined* clauses.

3.2 New Data

The annotated corpus prepared for the CODI-CRAC 2021 shared task consists of conversations from four well-known conversational datasets: the AMI corpus (Carletta, 2006), the LIGHT corpus (Urbanek et al., 2019), the PERSUASION corpus (Wang et al., 2019) and SWITCHBOARD (Godfrey et al., 1992). For each of these datasets, documents for about 15K tokens were annotated for development, and about the same number of tokens were annotated for testing, according to the ARRAU annotation scheme.

Switchboard SWITCHBOARD[8] (Godfrey et al., 1992) is one of the best known dialogue corpora. It consists of 1,155 five-minute spontaneous telephone conversations between two participants not previously acquainted with each other. In these conversations, callers question receivers on provided topics, such as child care, recycling, and news media. 440 speakers participate in these 1,155 conversations, producing 221,616 utterances. It was

[7] Identity anaphora also includes split antecedent plural anaphoric reference.

[8] https://catalog.ldc.upenn.edu/LDC97S62

annotated for dialogue acts by Stolcke et al. (1997)[9] and for information status by Nissim et al. (2004).

AMI The AMI corpus[10] (Carletta, 2006) is a collection of 100 hours of meeting recordings between several participants. The recordings include signals from close-talking and far-field microphones, individual and room-view video cameras, and output from a slide projector and an electronic whiteboard. Several types of annotation were carried out, including dialogue acts, topics, summaries, named entities, and focus of attention.

Light Amazon, Facebook, Google, and other AI companies have all created dialogue corpora in recent years to support their research on conversational agents. LIGHT (Urbanek et al., 2019) is one of the many recently created corpora available on the `Parl.ai` platform.[11] LIGHT is a large-scale fantasy text adventure game research platform for training agents that can both talk and act, interacting either with other models or with humans. The LIGHT corpus was entirely created through crowdsourcing at different levels. In the first round, workers created a number of settings (the King's palace, the dark forest, etc); then in a second round workers created fitting characters for each scenario, providing information about their background history, their personality, etc. Finally, in a third round, workers created dialogues between these characters.

Persuasion The Persuasion for Good corpus[12] (Wang et al., 2019) is a collection of online conversations generated by Amazon Mechanical Turk workers, where one participant (the persuader) tries to convince the other (the persuadee) to donate to a charity. 1017 conversations were collected in total, along with demographic data and responses to psychological surveys from users. Several speaker-level annotations were marked, including, e.g., demographics, the big five personality traits, etc.

3.3 Annotation

The dataset was annotated using the same MMAX2 tool (Müller and Strube, 2006) – indeed, almost exactly the same MMAX style – and by the same two annotators from the DALI team at Queen Mary University and University of Essex, Dr. Maris Camilleri and Dr. Paloma Carretero Garcia, who annotated and checked ARRAU Release 3, which is currently being prepared for release. However, due to time constraints, each document was only annotated by a single annotator, with spot checks carried out by the other annotator and Massimo Poesio (in ARRAU 3 each document was looked at by both annotators, and most documents were also independently checked by Massimo Poesio).

To prepare the data for the shared task, mentions were automatically extracted using the mention detector from Yu et al. (2020a), and the output converted into MMAX XML format.

3.4 The Corpus

Some basic statistics about the CODI-CRAC 2021 dataset are provided in Table 1. For each dataset, the Table reports number of documents, size in tokens, number of markables, and how many of these are Discourse Old (Identity Coreference) anaphors (DO), bridging references, and discourse deixis. With a total of 147,725 tokens and 41,807 markables, the CODI-CRAC 2021 dataset is to our knowledge the largest dataset annotated for anaphoric interpretation in dialogue, almost twice the size of ARRAU's `TRAINS` sub-corpus in tokens and more than twice its size in markables.

After annotation, the documents were converted into the CONLL-UA 'Extended' format used by the scorer, described by a document on the Universal Anaphora site.[13]

AMI, LIGHT and PERSUASION are freely available from the Shared Task Codalab site. ARRAU and SWITCHBOARD are distributed by LDC.[14]

4 Task Description

Following the structure of the CRAC 2018 Shared Task, CODI-CRAC 2021 was articulated around three tasks covering three key aspects of anaphoric interpretation: identity anaphora, bridging anaphora, and discourse deixis. Participants or groups could participate in one or more tasks.

[9] This version is available from `https://convokit.cornell.edu/documentation/switchboard.html`

[10] `https://groups.inf.ed.ac.uk/ami/corpus/`

[11] `https://parl.ai/projects/light/`

[12] `https://convokit.cornell.edu/documentation/persuasionforgood.html`

[13] `https://github.com/UniversalAnaphora/UniversalAnaphora/blob/main/UA_CONLL_U_Plus_proposal_v1.0.md`

[14] ARRAU is also freely available to any group that purchased the Penn Treebank and TRAINS-93 corpora from LDC.

		Docs	Tokens	Markables	DO	Bridging	Disc. Deix
AMI	dev	7	33741	8935	4400	850	230
	test	3	18260	4879	2300	633	118
LIGHT	dev	20	11495	3877	2120	381	62
	test	21	11824	3931	2174	415	80
PERSUASION	dev	22	9757	2929	1191	242	94
	test	28	12629	3839	1605	288	123
SWITCHBOARD	dev	11	14992	4025	1664	602	127
	test	22	35027	9392	3992	1190	263
Totals		134	147,725	41,807	19,446	4601	1097

Table 1: Statistics about the CODI-CRAC 2021 corpus (new datasets only)

4.1 Markable Settings

Bridging reference resolution and discourse deixis are very difficult tasks. In consideration of this, the Bridging (Task 2) and Discourse Deixis (Task 3) tasks were further divided into system and gold settings, according to whether the markables would be predicted by the system or provided by the organizers. The two settings were run in order – the gold setting became available after the runs under the system setting had been submitted. The two settings were scored separately and independently.

4.2 The Universal Anaphora Scorer

The new Universal Anaphora (UA) scorer was used to evaluate the systems. This is a Python scorer for the varieties of anaphoric reference covered by the Universal Anaphora guidelines, which include identity reference, split antecedent plurals, identification of non-referring expressions, bridging reference, and discourse deixis.

The scorer builds on the original Reference Coreference scorer[15] (Pradhan et al., 2014) developed for use in the CONLL 2011 and 2012 shared tasks on the ONTONOTES corpus (Pradhan et al., 2012) and its reimplementation in Python by Moosavi[16], which was already extended to evaluate non-referring expressions evaluation and cover singletons for the CRAC 2018 shared task (Poesio et al., 2018). The scorer reports scores for identity reference (with and without singletons and non-referring expressions – in the modality without singletons and non referring expressions the scorer is compatible with the original Coreference Reference scorer – split antecedents, bridging reference, and discourse deixis). For identity reference, the scorer reports the MUC, B^3, CEAF, CONLL (the unweighted average of MUC, B^3, and CEAF) (Pradhan et al., 2014), BLANC (Recasens and Hovy, 2011), and LEA (Moosavi and Strube, 2016) scores. The same scores are also computed for discourse deixis, which is treated as a generalized case of event coreference. For split antecedents, a generalization of these metrics following Paun et al. (2021) was developed. Entity F1 is computed for bridging–i.e., a system's interpretation is deemed correct as long as any mention of the correct anchor is found, as done e.g., in Hou et al. (2018).

4.3 Setting of the Scorer used in the Shared Task

The UA scorer can be run in a number of ways. The following settings were used for the individual tasks.[17]

Task 1 For Task 1, the Evaluating coreference relations (including split-antecedents) and singletons modality was used. Non-referring expressions identification were not scored.

```
python ua-scorer.py key system
```

Task 2 For Task 2, the scorer was called using the command:

```
python ua-scorer.py key system \
    keep_bridging
```

[15] https://github.com/conll/reference-coreference-scorers

[16] https://github.com/ns-moosavi/coval

[17] For a full description of the task(s), see https://github.com/sopankhosla/codi2021_scripts/blob/main/2021_CODI_CRAC_Introduction.md

Task 3 Finally, for Task 3, the scorer was called using the command:

```
python ua-scorer.py key system \
     evaluate_discourse_deixis
```

5 Submission Details

We used CodaLab to evaluate submissions and distribute the datasets. In the development phase, the participants only had access to out-of-domain training data (e.g., the ARRAU corpus) and in-domain validation data. They could submit results to the public leaderboard to evaluate their systems. In addition, we also released the scoring script on Github to reduce the dependency on CodaLab during model development. During the evaluation phase, we released the unseen test set across four sub-corpora. The submissions were evaluated on each sub-corpus individually and the final ranking for each task was performed by taking the mean of the four scores. Due to the lack of in-domain training data, the participants were allowed to use additional resources.

6 Baselines

We released one baseline system for each task. We derive the baselines for identity and bridging anaphora from current state-of-the-art methods, and set up a simple yet effective method for discourse deixis.

For Task 1, we used Xu and Choi (2020)'s coreference resolution model but without any higher-order inference.[18] We used the ONTONOTES (English) dataset for training and development. This model is then evaluated on CODI-CRAC 2021 datasets.

For Task 2, the baseline was derived from (Yu and Poesio, 2020). We used their single-task variant that is only trained on bridging annotations. We evaluated their best-performing model, which was trained on the RST sub-corpus of ARRAU, on CODI-CRAC 2021 data.[19]

The baseline for Task 3 leverages a simple heuristic that only considers demonstrative pronouns (*this*, *that*) as anaphors and considers the immediately preceding clause/utterance in the conversation to be their antecedent. Although simplistic, the algorithm achieves respectable scores on the CODI-CRAC 2021 development corpus. The performance of each baseline on different sub-corpora is shown in Tables 3, 4, and 5.

To help participants interested in building upon these baselines, we released the scripts used in the baseline pipeline to convert the CODI-CRAC 2021 data from the CONLL-UA format to JSON structure and vice-versa for each task.[20] The scripts contain modules that can also be used independently to transform conversations into a format compatible with different transformer-based encoders, allowing participants to set optional arguments to specify the *segment_size* and *tokenizer* for the conversion.

7 Participating Systems

A total of 55 individual participants registered for the CODI-CRAC 2021 shared task on CodaLab.[21] Among them, five teams submitted results for Task 1, three submitted results for Task 2, and two submitted results for Task 3. Teams UTD_NLP, KU_NLP, DFKI_TalkingRobots, Emory_NLP, and INRIA submitted system description papers. We summarize their approaches below (and in Table 2):

UTD_NLP participated in all three tasks. For identity anaphora, they deployed a pipeline architecture consisting of a mention detection component and an entity coreference component. The coreference component extends Xu and Choi (2020)'s implementation of Lee et al. (2018) by modifying the objective so that it can output singleton clusters, and enforces dialogue-specific constraints. They setup a similar architecture for discourse deixis. However, they slightly modified the objective function in Xu and Choi (2020) by classifying each span as a candidate anaphor, a candidate antecedent, or a non-mention in the mention detection stage, and resolving only candidate anaphors to candidate antecedents later. The team used a multi-pass sieve approach for bridging resolution to target same-head bridging links, with Yu and Poesio (2020)'s model as one of the sieves. In the gold setting, they trained an additional anaphor detection model (adapted from Yu and Poesio (2020)) to first identify the bridging anaphors from the gold markables before sending them to the sieves.

KU_NLP submitted results for tasks 1 and 2. For identity anaphora, they leveraged Cui and Zhang (2019)'s model with an ELECTRA-large backbone

[18] https://github.com/lxucs/coref-hoi/

[19] https://github.com/juntaoy/dali-bridging

[20] The necessary scripts are available from https://github.com/sopankhosla/codi2021_scripts

[21] Participants were allowed to create teams.

Track	Team	Baselines	Framework	Markable ID	Train. Data	Dev. Data
Anaphora Resolution	UTD_NLP	Xu and Choi (2020)	Pipeline of mention detection and entity coreference components. Modifies baseline to handle singleton clusters and enforce dialogue-specific constraints.	Adapted from Xu and Choi (2020)	CODI-CRAC 2021	CODI-CRAC 2021
	KU_NLP	-	(Cui and Zhang, 2019) with ELECTRA-large for mention detection. A pointer-network (Vinyals et al., 2015) based model for resolution.	Cui and Zhang (2019) + ELECTRA-large	CODI-CRAC 2021	CODI-CRAC 2021
	DFKI	-	(1) Workspace Coreference System incrementally clusters mentions using semantic similarity based on embeddings and lexical features; (2) Mention-to-Mention system pairs same entity mentions	SpaCy / BiLSTM-CRF	CODI-CRAC 2021	CODI-CRAC 2021
	Emory	Joshi et al. (2020)	Adapts the baseline to recognize singletons and encode speakers for all turns.	Joshi et al. (2020) + Singleton Recognition	CODI-CRAC 2021 + OntoNotes (Pradhan et al., 2012) + BOLT (Li et al., 2016)	CODI-CRAC 2021
Bridging Resolution	UTD_NLP	Yu and Poesio (2020)	A multi-pass sieve approach which used the baseline as one of the sieves and consists of a set of learning-based sieves (trained using SVMs) to target same-head bridging links.	Multi-task learning and SVM-based sieves	CODI-CRAC 2021	CODI-CRAC 2021
	KU_NLP	-	Solved as an MRC problem. Model takes query ("*What is related of ENTITY?*") and the conversation as input, and outputs the answer entity span.	Obtained as a part of the MRC system	CODI-CRAC 2021 + QUOREF	CODI-CRAC 2021
	INRIA	Joshi et al. (2019)	Leverage baseline's architecture to find the correct bridging antecedent in the gold setting.	Joshi et al. (2019)	CODI-CRAC 2021	CODI-CRAC 2021
Discourse Deixis Resolution	UTD_NLP	Xu and Choi (2020)	Builds upon the baseline by classifying each span into candidate anaphor, a candidate antecedent, or a non-mention. Only resolves candidate anaphors to candidate antecedents.	Obtained as part of joint mention detection and deixis resolution	CODI-CRAC 2021	CODI-CRAC 2021
	DFKI	-	A Siamese Network to detect discourse anaphor-antecedent pairs.	SpaCy	CODI-CRAC 2021	CODI-CRAC 2021

Table 2: Summary of the Participating Systems

Team	LIGHT	AMI	PERS.	SWBD.	Avg.
Eval AR					
Emory	**80.33 (1)**	**63.98 (1)**	**78.41 (1)**	**74.49 (1)**	**74.3 (1)**
UTD_NLP	79.56 (2)	57.38 (3)	77.50 (2)	72.64 (2)	71.8 (2)
KU_NLP	69.16 (3)	57.59 (2)	71.09 (3)	65.67 (3)	65.9 (3)
DFKI (1)	64.99 (4)	43.93 (4)	59.93 (4)	53.55 (4)	55.6 (4)
SCIR[22]	55.92 (6)	39.46 (5)	52.25 (6)	51.63 (5)	49.8 (5)
Baseline	52.45 (7)	36.11 (6)	51.97 (7)	45.80 (7)	46.6 (6)
DFKI (2)	61.26 (5)	00.00 (7)	59.20 (5)	51.24 (6)	42.9 (7)

Table 3: Performance on Task 1 (Evaluation Phase) – Identity Anaphora (CoNLL Avg. F1)

Team	LIGHT	AMI	PERS.	SWBD.	Avg.
Eval Br (Gold)					
UTD_NLP	**19.73**	**19.65**	**31.40**	**21.10**	**23.0**
KU_NLP	16.67	15.30	18.79	18.33	17.3
INRIA	9.35	6.00	16.28	7.79	9.9
Baseline	6.35	6.21	13.77	5.39	7.9
Eval Br (Pred)					
UTD_NLP	**13.98**	**13.33**	**21.92**	**15.26**	**16.1**
KU_NLP	13.46	10.25	12.32	10.99	11.8
Baseline	6.01	4.94	9.34	3.78	6.0

Table 4: Performance on Task 2 (Evaluation Phase) – Bridging Anaphora (Entity F1)

Team	LIGHT	AMI	PERS.	SWBD.	Avg.
Eval DD (Gold)					
UTD_NLP	**43.44**	**36.91**	**52.09**	**40.44**	**43.2**
Eval DD (Pred)					
UTD_NLP	**42.70**	**35.35**	**39.64**	**35.43**	**38.3**
DFKI	20.97	17.43	23.76	23.86	21.5
Baseline	12.12	15.75	18.27	13.55	14.9

Table 5: Performance on Task 3 (Evaluation Phase) – Discourse Deixis (CoNLL Avg. F1)

for mention detection. The resulting mention representation, created from the constituent token representations, is then fed to a pointer-network (Vinyals et al., 2015) based coreference resolution model for clustering. They solved the bridging resolution problem using a machine reading comprehension framework, where they constructed a query for each entity of the form – *"What is related of ENTITY?"*. The input of their model is the query and the document (i.e., all utterances of dialogue), and the output is the entity span that is the answer for the query.

DFKI_TalkingRobots (DFKI) put forward two systems as their final submissions for the identity anaphora task. The Workspace Coreference System (WCS) attempts to incrementally cluster mentions using semantic similarity based on embeddings combined with lexical feature heuristics, whereas their Mention-to-Mention (M2M) architecture only focuses on mention pairs to make the decision. WCS and M2M use SpaCy and BiLSTM-CRF based mention detection components, respectively. For discourse deixis, the team deployed a Siamese Network to detect discourse anaphor-antecedent pairs. They only focused on demonstrative pronouns (as anaphors), which they detected using the SpaCy NLP pipeline.

Emory_NLP (Emory) only participated in the identity anaphora task. Their system was adapted from the end-to-end neural coreference resolution model of Joshi et al. (2020). They recognized singletons, encoded speakers for all turns, and leveraged other out-of-domain datasets during training.

INRIA submitted an end-to-end transformer-based model fine-tuned for the bridging resolution task. They formulated the bridging problem as antecedent selection, and leveraged Lee et al. (2018); Joshi et al. (2019)'s architecture to find the correct antecedent.

8 Results and Discussion

In this section, we report the results of all systems submitted for each task and discuss the differences among these approaches.

[22]Team SCIR did not submit a system description paper.

8.1 Task 1 – Identity Anaphora

Task 1 saw the highest interest as five teams submitted a total of 36 runs to the official leaderboard. As discussed earlier, we report the CoNLL Avg. F1 score for each sub-corpus separately.

The results on different sub-corpora are reported in Table 3. All five runs outperform the baseline in terms of the CoNLL Avg. F1 score. The best run on all four sub-corpora was submitted by *lxucs* (the Emory team) achieving around 75% or more CoNLL Avg. F1 on LIGHT, PERSUASION, and SWITCHBOARD, and the rankings across the four datasets remained relatively stable. AMI proved to be the toughest sub-corpus with the highest score being around 64%, which is 10 percentage points below that of SWITCHBOARD. This is in line with the organizers' expectations, as AMI proved the most difficult corpus to annotate, with lots of uncertainty and ambiguity. Also, the conversations in AMI are substantially longer than the other three datasets and hence require systems to consider long-distance relationships between mentions. Despite the differences in absolute performance on each sub-corpus, the best-performing system improved over the baseline by an impressive 30 CoNLL Avg. F1 percentage points.

8.2 Task 2 – Bridging Anaphora

Three teams participated in Task 2 with *INRIA* only participating in the gold mention setting. We report Entity F1 scores for each sub-corpora and calculate rankings based on the mean score across the four datasets.

Table 4 summarizes the performance of each run. For the predicted setting Eval-Br (Pred), where the systems need to take a raw conversation as input during the inference time and perform both markable identification and bridging resolution, *UTD_NLP* performs the best across all four sub-corpora. Although both participating systems perform similarly across SWITCHBOARD, AMI, and LIGHT (10 – 15 Entity F1 points), *UTD_NLP* achieves an unusually high score on the PERSUASION dataset (for this task), reaching an Entity F1 of 21.92 percentage points. We see a similar trend in the gold setting (Table 2), where *UTD_NLP* and *INRIA* score 8–10 absolute percentage points higher on PERSUASION (31.40 and 16.28 respectively) as compared to their scores on the other three datasets. Finally, the performance of the runs in the gold setting is substantially higher than that in the predicted setting for both *UTD_NLP* and *KU_NLP* even though the gold setting does not distinguish between the markables that are relevant for the three tasks in the competition.

8.3 Task 3 – Discourse Deixis

We received 25 runs for Task 3. Two teams (*UTD_NLP*, *DFKI_TalkingRobots*) submitted to the predicted mention setting with *UTD_NLP* achieving performance around 35–42 CoNLL Avg. F1 percentage points on the different sub-corpora, almost doubling the score of the second team. For the gold setting, where we released the gold markables, the system submitted by *UTD_NLP* did not improve their scores substantially on LIGHT, AMI, or SWITCHBOARD from the predicted setting. However, they managed a jump of more than 12 CoNLL Avg. F1 points on PERSUASION. All teams outperform the baseline system on all sub-corpora. The performance of both teams on Task 3 is summarized in Table 5.

8.4 Discussion

In retrospect, organizing this shared task required tackling a few too many issues in a short time, from annotating the data to developing a new scorer to devising fair ways to use the scorer to assess systems, to be able to address all of them in a completely satisfactory way. This is why we decided to run the same task again next year, with new test data but maintaining the same genres and annotation guidelines. The ARRAU 3 data and annotation manual should also be available, addressing some of the concerns raised by the participants.

It would be premature to infer too much about the tasks or the genres on the basis of this first experience, but some interesting issues are already emerging from the analysis papers, such as the complex use of first and second person pronouns in these datasets. We do hope it will be possible to carry out a more extensive analysis of the differences between these dialogue datasets and datasets based on written text and annotated for the same phenomena, such as the RST subcorpus of ARRAU.

We would also like to congratulate all participants for gallantly tackling these relatively new tasks and new datasets yet generally outperforming the baselines, some of which were non-trivial.

9 Conclusion and Future Work

We presented a general overview of the CODI-CRAC 2021 shared task. As the first instance in this series, CODI-CRAC 2021 focused on resolving three types of anaphoric relations in dialogues: identity, bridging references, and discourse deixis. In addition, we described the CODI-CRAC 2021 corpus, which contains sub-corpora from different conversation genres newly annotated for the above-mentioned relations. While the teams were encouraged to create systems with generalizable representations using other out-of-domain state-of-the-art coreference datasets during training, only the best-performing team for the entity coreference track did so. Finally, we included a brief summary of the different approaches used by the participants to tackle different tasks within the shared-task.

For the shared-task's next installment, we plan to release the full annotation guidelines to reduce the ambiguity about annotation principles. Based on suggestions from participating teams (Team *DFKI*), we will introduce separate tracks to fairly evaluate the systems trained on the provided data vs. systems (pre-)trained on additional data, and discuss the possibility of using more fine-grained evaluation metrics to differentially evaluate cases of coreference based on their difficulty levels. (E.g., *a black cat* and *the black cat* are much easier to resolve compared to *a black cat* and *the dark and furry creature*.)

Acknowledgments

We are very grateful to Maris Camilleri and Paloma Carretero Garcia, who annotated the dataset very quickly, and to the Linguistic Data Consortium, who very generously made the ARRAU and SWITCHBOARD data available to the participants to the competition. We would also like to thank Team *DFKI* for their suggestions for the future of the shared task.

Massimo Poesio wishes to thank Sharid Loáiciga, Anja Nedoluzhko, Arndt Riester, Amir Zeldes, and several other participants in the Universal Anaphora initiative for discussions about anaphoric annotation and anaphoric annotation in dialogue in particular. The work of Massimo Poesio and Juntao Yu was funded by the the DALI project, ERC Grant 695662. The annotation was funded in part by DALI, in part by HITS Heidelberg.

References

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).

Donna Byron. 2002. Resolving pronominal references to abstract entities. In *Proc. of the ACL*, pages 80–87.

Donna Byron and James Allen. 1998. Resolving demonstrative anaphora in the trains-93 corpus. In *Proc. of the Second Colloquium on Discourse, Anaphora and Reference Resolution*. University of Lancaster.

Jean Carletta. 2006. Announcing the AMI meeting corpus. *The ELRA Newsletter*, 11(1):3–5.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In J. Kuppevelt and R. Smith, editors, *Current Directions in Discourse and Dialogue*, pages 85–112. Kluwer.

Wallace L. Chafe. 1980. *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production*. Ablex, Norwood, NJ.

Herbert H. Clark. 1977. Bridging. In P. N. Johnson-Laird and P.C. Wason, editors, *Thinking: Readings in Cognitive Science*, pages 411–420. Cambridge University Press, London and New York.

Kevin Bretonnel Cohen, Arrick Lanfranchi, Miji Joo-young Choi, Michael Bada, William A. Baumgartner Jr., 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(372).

Leyang Cui and Yue Zhang. 2019. Hierarchically-refined label attention network for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4115–4128.

Miriam Eckert and Michael Strube. 2001. Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*.

Carola Eschenbach, Christopher Habel, Michael Herweg, and Klaus Rehkämper. 1989. Remarks on plural anaphora. In *Proceedings of the fourth conference on European chapter of the Association for Computational Linguistics*, pages 161–167. Association for Computational Linguistics.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: telephone speech corpus for research and development . acoustics,. In *IEEE International Conference on Speech, and Signal Processing, ICASSP-92*, volume 1, pages 517–520.

Loïc Grobol. 2020. *Coreference resolution for spoken French*. Ph.D. thesis, Université Sorbonne Nouvelle.

John A. Hawkins. 1978. *Definiteness and Indefiniteness*. Croom Helm, London.

Yufang Hou. 2020. Bridging anaphora resolution as question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1428–1438, Online. Association for Computational Linguistics.

Yufang Hou, Katja Markert, and Michael Strube. 2018. Unrestricted bridging resolution. *Computational Linguistics*, 44(2):237–284.

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.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5803–5808, Hong Kong, China. Association for Computational Linguistics.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic*. D. Reidel, Dordrecht.

Ben Kantor and Amir Globerson. 2019. Coreference resolution with entity equalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 673–677, Florence, Italy. Association for Computational Linguistics.

Hideo Kobayashi and Vincent Ng. 2020. Bridging resolution: A survey of the state of the art. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3708–3721, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hideo Kobayashi and Vincent Ng. 2021. Bridging resolution: Making sense of the state of the art. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1652–1659, Online. Association for Computational Linguistics.

Varada Kolhatkar and Graeme Hirst. 2014. Resolving shell nouns. In *Proc. of EMNLP*, pages 499–510, Doha, Qatar.

Varada Kolhatkar, Adam Roussel, Stefanie Dipper, and Heike Zinsmeister. 2018. Anaphora with non-nominal antecedents in computational linguistics: a Survey. *Computational Linguistics*, 44(3):547–612.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692.

Xuansong Li, Martha Palmer, Nianwen Xue, Lance Ramshaw, Mohamed Maamouri, Ann Bies, Kathryn Conger, Stephen Grimes, and Stephanie Strassel. 2016. Large multi-lingual, multi-level and multi-genre annotation corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 906–913, Portorož, Slovenia. European Language Resources Association (ELRA).

Quan Liu, Hui Jiang, Andrew Evdokimov, Zhen-Hua Ling, Xiaodan Zhu, Si Wei, and Yu Hu. 2017. Cause-effect knowledge acquisition and neural association model for solving a set of winograd schema problems. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2344–2350.

Pengcheng Lu and Massimo Poesio. 2021. Coreference resolution for the biomedical domain: A survey. In *Proc. of the CRAC Workshop*.

Ana Marasović, Leo Born, Juri Opitz, and Anette Frank. 2017. A mention-ranking model for abstract anaphora resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 221–232, Copenhagen, Denmark. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Katja Markert, Yufang Hou, and Michael Strube. 2012. Collective classification for fine-grained information status. In *Proc. of the ACL*, Juju island, Korea.

Natalia N. Modjeska. 2003. *Resolving other anaphors*. Ph.D. thesis, University of Edinburgh.

Nafise S. Moosavi and Michael Strube. 2016. A proposal for a link-based entity aware metric. In *Proc. of ACL*, pages 632–642, Berlin.

Mark-Christoph Müller. 2008. *Fully Automatic Resolution of It, This And That in Unrestricted Multy-Party Dialog*. Ph.D. thesis, Universität Tübingen.

Mark-Christoph Müller and Michael Strube. 2006. Multi-level annotation of linguistic data with mmax2. In S. Braun, K. Kohn, and J. Mukherjee, editors, *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods*, volume 3 of *English Corpus Linguistics*, pages 197–214. Peter Lang.

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*.

Costanza Navarretta. 2000. Abstract anaphora resolution in Danish. In *Proc. of the 1st SIGdial Workshop on Discourse and Dialogue*, pages 56–65. ACL.

Anna Nedoluzhko. 2013. Generic noun phrases and annotation of coreference and bridging relations in the prague dependency treebank. In *Proc. of LAW*, pages 103–111.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, and Daniel Zeman. 2021. Coreference meets universal dependencies – a pilot experiment on harmonizing coreference datasets for 11 languages. ÚFAL Technical Report TR-2021-66, Charles University, Prague.

Malvina Nissim, Shipra Dingare, Jean Carletta, and Mark Steedman. 2004. An annotation scheme for information status in dialogue. In *Proc. of LREC*.

Silviu Paun, Juntao Yu, Nafise Moosavi, and Massimo Poesio. 2021. Scoring coreference chains with split-antecedent anaphors and other entities constructed from a discourse model. Submitted.

Massimo Poesio. 2004. Discourse annotation and semantic annotation in the GNOME corpus. In *Proc. of the ACL Workshop on Discourse Annotation*, pages 72–79, Barcelona.

Massimo Poesio. 2016. Linguistic and cognitive evidence about anaphora. In M. Poesio, R. Stuckardt, and Y. Versley, editors, *Anaphora Resolution: Algorithms, Resources and Applications*, chapter 2. Springer.

Massimo Poesio and Ron Artstein. 2008. Anaphoric annotation in the ARRAU corpus. In *Proc. of LREC*, Marrakesh.

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, 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 *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 11–22, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Massimo Poesio and Renata Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35, Baltimore, Maryland. Association for Computational Linguistics.

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.

Sameer S. Pradhan, Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla. 2007. Unrestricted coreference: Indentifying entities and events in ontonotes. In *Proc. IEEE International Conference on Semantic Computing (ICSC)*, Irvine, CA.

Ellen F. Prince. 1981. Toward a taxonomy of given-new information. In P. Cole, editor, *Radical Pragmatics*, pages 223–256. Academic Press, New York.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The Winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789, Jeju Island, Korea. Association for Computational Linguistics.

Marta Recasens and Ed Hovy. 2011. Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*.

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

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:8732–8740.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van-Ess-Dykema, and Marie Meteer. 1997. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–371.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, Samuel Humeau, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. ArXiv preprint arXiv:1903.03094.

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*.

Hardik Vala, Andrew Piper, and Derek Ruths. 2016. The more antecedents, the merrier: Resolving multi-antecedent anaphors. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2287–2296, Berlin, Germany. Association for Computational Linguistics.

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

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. *Advances in Neural Information Processing Systems*, 28:2692–2700.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proc. of ACL*.

Bonnie L. Webber. 1991. Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6(2):107–135.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Sam Wiseman, Alexander M. Rush, Stuart Shieber, and Jason Weston. 2015. Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1416–1426, Beijing, China. Association for Computational Linguistics.

Liyan Xu and Jinho D Choi. 2020. Revealing the myth of higher-order inference in coreference resolution. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8527–8533.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020a. Neural mention detection. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*.

Juntao Yu, Nafise Sadat Moosavi, Silviu Paun, and Massimo Poesio. 2020b. Free the plural: Unrestricted split-antecedent anaphora resolution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6113–6125, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Juntao Yu, Nafise Sadat Moosavi, Silviu Paun, and Massimo Poesio. 2021. Stay together: A system for single and split-antecedent anaphora resolution. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Juntao Yu and Massimo Poesio. 2020. Multitask learning based neural bridging reference resolution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3534–3546, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Ethan Zhou and Jinho D. Choi. 2018. They exist! introducing plural mentions to coreference resolution and entity linking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 24–34, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yilun Zhu, Sameer Pradhan, and Amir Zeldes. 2021. OntoGUM: Evaluating contextualized SOTA coreference resolution on 12 more genres. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 461–467, Online. Association for Computational Linguistics.