The CODI-CRAC 2022 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue

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

Abstract

The CODI-CRAC 2022 Shared Task on Anaphora Resolution in Dialogues is the second edition of an initiative focused on detecting different types of anaphoric relations in conversations of different kinds. 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. The second edition of the shared task maintained the focus on these relations and used the same datasets as in 2021, but new test data were annotated, the 2021 data were checked, and new subtasks were added. In this paper, we discuss the annotation schemes, the datasets, 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 230 runs from three 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, and on genres other than news.

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 also 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 and mentions jointly created across utterances (Poesio and Rieser, 2010) or whose function is to establish common ground rather than refer (Clark and Brennan, 1990; Heeman and Hirst, 1995). 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 normally absent from news or Wikipedia articles, which constitute the bulk of current datasets for coreference resolution (Poesio et al., 2016). 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.

*Work was done prior to joining AWS AI Labs.

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 CODI-CRAC 2021 Shared Task in Anaphora Resolution in Dialogue (Khosla et al., 2021) was organized to address this need for datasets about anaphoric reference in dialogue by providing participants with the opportunity to develop automated approaches for anaphora resolution that tackle less studied forms of anaphora as well as coreference, and generalize to different types of conversational setups. A number of groups participated to this first edition, but we organizers also realised that the community could benefit from a second edition using more data and more cleaned-up, adding more tasks, and improving the evaluation. As a result, we organized this year's second edition. [1] Like the first edition, CODI-CRAC 2022 involved three tasks that individually tackle a particular anaphoric relation: identity, bridging, and discourse deixis, in four conversational datasets from different domains newly annotated with the above-mentioned relations. Unlike the first edition, participants also had training data in those four domains, in addition to development and test sets. To accommodate for systems that use gold/predicted mentions for bridging and discourse deixis tasks, we set up separate leaderboards for the two settings.

In this paper wepresent an overview of the CODI-CRAC 2022 shared task. We begin by providing some background in Section 2 and introducing the new CODI-CRAC 2022 corpus in Section 3. We then provide an extensive overview of the different CODI-CRAC 2022 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.

[1] https://codalab.lisn.upsaclay.fr/competitions/614

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. The CODI-CRAC 2021 and 2022 Shared Tasks covered the range of anaphoric relations included in the first Universal Anaphora survey of phenomena to be covered (see below)

Split-antecedent anaphora **Split-antecedent anaphors** (Eschenbach et al., 1989; Kamp and Reyle, 1993) are cases of plural identity reference to sets composed of two or more entities introduced by separate noun phrases, as in (2).

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

Such references are annotated in, e.g., ARRAU (Uryupina et al., 2020), GUM (Zeldes, 2017) and *Phrase Detectives* (Poesio et al., 2019).

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., *[this]* in the same example, which 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); 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 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 (4), where bridging reference / associative anaphora *the roof* refers to an object which is related to / associated with, but not identical to, the *hall*.

(4) 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]].

2.2 Universal Anaphora

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 TÜBA-DZ (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).

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 was used in CODI-CRAC 2021 and for this shared task (Yu et al., 2022).

2.3 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 2022 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 2022 Corpus

One of the objectives of the CODI-CRAC shared tasks 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 made available as training material. In addition, new data

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

[3] https://universaldependencies.org/

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 Penn Treebank (Marcus et al., 1993). The TRAINS domain of task-oriented dialogues includes a complete annotation of the TRAINS-93 corpus[5] and the pilot dialogues in the so-called TRAINS-91 corpus. In addition, ARRAU includes a complete annotation of the spoken narratives in the Pear Stories (Chafe, 1980), and documents in the medical and art history genres from the GNOME corpus (Poesio, 2004).

Annotation scheme Following the CRAC 2018 shared task, a revised version of the annotation guidelines was produced, as part of the work on the ARRAU 3 release of the corpus. The new annotation guidelines were completed after CODI-CRAC 2021 and made available on the corpus page.[6] The new guidelines were used in CODI-CRAC 2022 to check the annotation of the documents already annotated for CODI-CRAC 2021 and to annotate new data. For more information on the scheme, please consult the manual or, for a quick summary, (Khosla et al., 2021).

3.2 New Data

The annotated corpus created for CODI-CRAC 2022 consists of conversations from the same well-known conversational datasets already used in CODI-CRAC 2021: 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 in 2021 for development according to the ARRAU annotation scheme, and about the same number of tokens were annotated for testing. For this year's shared task, the development data from 2021 were used as training data; the test data from 2021 were used as development data; and new test data were annotated.

Switchboard SWITCHBOARD[7] (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 annotated for dialogue acts by Stolcke et al. (1997)[8] and for information status by Nissim et al. (2004).

AMI The AMI corpus[9] (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.[10] 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.

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

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

[6] https://github.com/arrauproject/data/blob/main/ARRAU_3_Annotation_Manual_1.0.pdf

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

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

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

[10] https://parl.ai/projects/light/

Persuasion The Persuasion for Good corpus[11] (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 – used to annotate and check ARRAU Release 2 and Release 3. But this time, the annotation work was divided between the DALI team at Queen Mary University (Maris Camilleri and Paloma Carretero Garcia, who have been annotating ARRAU 3), and a team at CMU coordinated by Lori Levin (Taiqi He and Katherine Zhang). This division of labor made it possible to (i) ensure that every new document would be annotated by at least two annotators, (ii) re-check the documents already annotated in 2021, and (iii) test the reliability of the scheme.

3.4 The Corpus

Some basic statistics about the CODI-CRAC 2022 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 214,625 tokens and 60,993, the CODI-CRAC 2022 dataset is to our knowledge the largest dataset annotated for anaphoric interpretation in dialogue. It is also one of the largest datasets annotated for bridging references.

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.[12]

AMI, LIGHT and PERSUASION are freely available from the Shared Task Codalab site. ARRAU and SWITCHBOARD are distributed by LDC.[13]

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

[12] https://github.com/UniversalAnaphora/UniversalAnaphora/blob/main/documents/UA_CONLL_U_Plus_proposal_v1.0.md

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

4 Task Description

Following the structure of the last year's Shared Task, CODI-CRAC 2022 covers three key aspects of anaphoric interpretation: identity anaphora, bridging anaphora, and discourse deixis. Participants or groups could participate in one or more tasks.

4.1 Markable Settings

To address the challenge of the bridging reference resolution and discourse deixis tasks, in addition to the predicted (Pred) and gold mention (Gold M) settings from last year, a gold anaphors (Gold A) setting is added to those tasks. In total, the Bridging (Task2) and Discourse Deixis (Task 3) tasks have three settings: Pred: the system is responsible for predicting their mentions; Gold M: with the gold mentions provided and Gold A: both gold anaphors and gold mentions were provided. The three settings were run in the order of Pred, Gold M and Gold A – the later settings became available after the runs under the previous settings had been submitted. The three settings were scored separately and independently.

4.2 Evaluation Settings

Same as last year, the Universal Anaphora (UA) scorer (Yu et al., 2022; Paun et al., 2022) was used to evaluate the systems. The same settings for last year's shared task were used, more specifically, the settings for the individual tasks are as follows:[14]

Task 1 For Task 1, we use the default settings of the scorer where the identity relations (including split-antecedents) and singletons were evaluated. Non-referring expressions were excluded from the evaluation.

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

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

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

Task 3 Finally, for Task 3, the scorer was called using the following command.

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

[14] For a full description of the task(s), see https://github.com/juntaoy/codi-crac2022_scripts/blob/main/2022_CODI_CRAC_Introduction.md

		Docs	Tokens	Markables	DO	Bridging	Disc. Deix
LIGHT	train	20	11495	3907	2132	381	72
	dev	21	11824	3941	2181	424	84
	test	38	22017	7330	3770	812	128
AMI	train	7	33741	8918	4579	853	230
	dev	3	18260	4870	2350	638	118
	test	3	16562	3990	2007	432	118
PERSUASION	train	21	9185	2743	1242	248	95
	dev	27	12198	3697	1715	316	133
	test	33	14719	4233	2111	304	105
SWITCHBOARD	train	11	14992	4024	1679	589	128
	dev	22	35027	9392	3991	1165	265
	test	12	14605	3888	1606	464	107
Total		218	214625	60933	29363	6626	1583

Table 1: Statistics about the CODI-CRAC 2022 corpus (new datasets only)

5 Submission Details

The shared task was hosted on a single CodaLab page, including evaluations and datasets distribution. The competition consists of three development phases and seven evaluation phases. In the development phases, a small in-domain training set for each domain alongside a large out-of-domain training set (i.e. the ARRAU corpus) is available. In addition, a validation set for each domain is also provided. The development phases are handy tools to get the systems prepared for the evaluation phases. Apart from the development phases, the participants can also download the scoring script to evaluate their systems offline. During the evaluation phases, the different versions of the unseen test sets (Pred, Gold M, Gold A) were released incrementally to accommodate the needs of the evaluation phases. The submissions were evaluated individually on each of the four domains, and then the macro-average of the four scores are used for the final ranking of individual tasks. Apart from the corpora provided by us, additional resources were also permitted.

6 Baselines

We used the same baseline systems from last year's shared task, and further, evaluate those baselines in the newly introduced phases. More precisely the baselines for identity anaphora and bridging reference resolution tasks are derived from state-of-the-art neural models, whereas the discourse deixis baseline is a simple but effective system based on heuristic rules.

For identity anaphora resolution (Task 1), we used the coreference resolution model provided by the Xu and Choi (2020)[15]. More specifically, we use their SpanBERT setting without any higher-order inference (SpanBERT + no HOI), The model was trained with the ONTONOTES (English) dataset and then evaluated directly on CODI-CRAC 2022 datasets without fine-tuning.

For bridging reference resolution (Task 2), we use the single-task variant of the Yu and Poesio (2020) system[16]. The system is trained on the bridging annotations of the RST sub-corpus of ARRAU. Since the system do not predict the mentions itself, for the predicted mention setting (Pred), we supply the system with mentions predicted by Yu et al. (2020a)'s mention detector (BIAFFINE MD)[17]. The mention detector was also trained on the same RST sub-corpus of ARRAU. For Gold M and Gold A settings, we use the gold mentions and anaphoras provided respectively. The system is evaluated on CODI-CRAC 2022 data without further training.

For discourse deixis (Task 3), the baseline for predicted mention setting (Pred) uses two simple heuristics: first only considers demonstrative pronouns (*this*, *that*) as anaphors and then uses the immediately preceding clause/utterance in the conver-

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

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

[17] https://github.com/juntaoy/dali-md

sation to be their antecedent. For the gold mention setting (Gold M) we further restrict the anaphors to be the intersection of the demonstrative pronouns and the gold mentions and then apply the same rule for antecedent selection. For the gold anaphor setting (Gold A), the baseline links the gold anaphors to their immediately preceding clause/utterance. The heuristic-based baselines are then evaluated on the CODI-CRAC 2022 data of all four domains.

The performance of our baselines on different sub-corpora is shown in Tables 3, 4, and 5 alongside the participant systems.

A helper script developed from last year's shared task is available to help participants convert the CONLL-UA format to and back from the various JSON format used by our baselines[18].

7 Participating Systems

Similar to last year, a total of 54 individual participants registered for the CODI-CRAC 2022 shared task on CodaLab. Among them, three teams submitted results for Task 1, and two submitted results for Task 2 and Task 3. Apart from Emory_NLP, all the teams from last year participated in this year's shared task, but DFKI and INRIA joined forces to participate as one team. All three teams (UTD_NLP, KU_NLP, DFKI-INRIA) submitted system description papers. We summarize their approaches below and in Table 2.

UTD_NLP participated in all three tasks. For identity anaphora, the authors built a pipeline system consisting of three components: a mention detector, an entity coreference resolver and a non-referring/entity classifier. All three components use the same underlining system they used in last year's shared task (Kobayashi et al., 2021), a multi-task learning approach adapted from the Xu and Choi (2020) system for mention detector and coreference resolution. The training objectives and priorities, however, were configured differently to maximise the performance of the individual tasks. Finally, those components were used in a pipeline fashion to deliver their final results. For discourse deixis, a system similar to Xu and Choi (2020)'s was used. They use both heuristics and a binary classifier to supply the anaphors. For each anaphor, antecedents were selected from up to 10 immediate previous utterances. The team based their bridging resolution system on the Yu and Poesio (2020)'s model, with additional dialogue-specific features included. The main focus of this year was on exploring the different pre-training and fine-tuning strategy. In total, four different training strategies were evaluated by them..

KU_NLP submitted results for identity anaphora resolution (task 1). The team proposed a pipeline system that resolves the mentions separately from the coreference resolution. The mention detection part solves the problem by classifying all possible mentions into mentions and non-mentions. The predicted mentions then feed into the coreference part of the system that solves the task in a mention-pair fashion. Additional speaker features were used to leverage the mention representations.

DFKI-INRIA participated in all three tasks. For the identity anaphora task, they utilise the Workspace Coreference System (WCS) (Anikina et al., 2021) they introduced in last year's shared task with the Xu and Choi (2020) system. The singletons predicted by the WCS system are added to the Xu and Choi (2020) to create their final results. Similar to the WCS system, the mentions are predicted separately using SpaCy. For bridging, they build their system on a simplified Joshi et al. (2019) system with mention pruning and coarse-to-fine steps removed. They only submitted to the Gold A phase, where gold mentions and gold anaphors were provided. For discourse deixis, the team employ a multi-task learning approach based on the Xu and Choi (2020) system, the system first uses heuristics to find the candidate anaphors, then resolve the antecedents and finally uses an anaphora type classifier to filter out the identity, non-referring anaphors. The system also used several linguistic features (e.g. PoS, dependency relations) to aid the anaphora type classification.

8 Results and Discussion

8.1 Task 1 – Identity Anaphora

All three teams participated the task 1, in total they made 55 runs to the official leaderboard. For this task, we report the CoNLL average F1 scores for each sub-corpus and take the macro-average of them to rank the participating systems.

As shown in Table 3, all the participating systems outperform the baseline by large margins (up to 27% on the macro-average scores). The best result was achieved by the UTD_NLP team, with large improvements over the baseline by more than

[18] https://github.com/juntaoy/codi-crac2022_scripts

Track	Team	Baselines	Framework	Markable ID	Train. Data	Dev. Data
Anaphora Resolution	UTD_NLP	Xu and Choi (2020)	A pipeline of mention detection, entity coreference and non-referring/mention removal components. Modifies baseline to handle singleton clusters and enforce dialogue-specific constraints.	Adapted from Xu and Choi (2020)	CODI-CRAC 2022 + OntoNotes	CODI-CRAC 2022
	KU_NLP	-	A pipeline system that predicts the mentions and resolves the coreference separately.	Span classification	CODI-CRAC 2022	CODI-CRAC 2022
	DFKI-INRIA	Xu and Choi (2020), Anikina et al. (2021)	The Xu and Choi (2020) was used as the main system for coreference and the output is supplemented with singletons from the Anikina et al. (2021) system	SpaCy	CODI-CRAC 2022 + OntoNotes	CODI-CRAC 2022
Bridging Resolution	UTD_NLP	Yu and Poesio (2020)	Build upon the baseline with SpanBERT as the backbone. Additional dialogue-specific features were used.	Adapted from Xu and Choi (2020)	CODI-CRAC 2022	CODI-CRAC 2022
	DFKI-INRIA	Joshi et al. (2019)	Remove the coarse-to-fine score of the baseline and resolve the bridging in the Gold A setting.	Joshi et al. (2019)	CODI-CRAC 2022 + BASHI + ISNotes	CODI-CRAC 2022
Discourse Deixis Resolution	UTD_NLP	Xu and Choi (2020)	Using heuristic and a binary classifier to select candidate anaphors. For each selected anaphor up to 10 previous utterances were used as candidate antecedents. Then the system assigns antecedents to each of the candidate anaphors	Obtained as part of joint mention detection and deixis resolution	CODI-CRAC 2022	CODI-CRAC 2022
	DFKI-INRIA	Xu and Choi (2020)	A multi-task learning system learning on both coreference and discourse deixis. With additional anaphor type classifier to filter non-discourse deixis anaphors.	Heuristic for anaphors; antecedents were predicted by the baseline	CODI-CRAC 2022	CODI-CRAC 2022

Table 2: Summary of the Participating Systems

Team	LIGHT	AMI	PERS.	SWBD.	Avg.
Eval AR					
UTD_NLP	**82.23**	**62.90**	**79.20**	**75.81**	**75.04**
DFKI-INRIA	72.06	51.41	69.87	60.61	63.49
KU_NLP	68.27	48.87	69.06	60.99	61.80
Baseline	54.23	34.14	53.16	49.30	47.71

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

Team	LIGHT	AMI	PERS.	SWBD.	Avg.
Eval Br (Gold A)					
UTD_NLP	**46.80**	**39.35**	**56.91**	**44.40**	**46.87**
DFKI-INRIA	37.68	35.23	50.99	35.78	39.92
Baseline	29.93	22.69	37.83	30.39	30.21
Eval Br (Gold M)					
UTD_NLP	**26.77**	**19.65**	**34.59**	**22.74**	**25.94**
Baseline	4.99	8.77	11.49	7.08	8.08
Eval Br (Pred)					
UTD_NLP	**23.25**	**13.42**	**27.75**	**19.72**	**21.04**
Baseline	4.01	4.66	8.45	4.00	5.28

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

25% for all four sub-corpora. For LIGHT and PERSUASION, the system achieved CoNLL Avg. F1 scores of 80% or more, the result on the SWITCHBOARD followed closely with an F1 of 76%. The system performance on the toughest sub-corpus (AMI) is way below the other sub-corpora a large 20% gap between LIGHT and AMI are visible across all the participant system as well as the baseline. The reason leads to the large gaps in performance between AMI and other sub-corpora is mainly due to the conversations in AMI being substantially longer than the other corpora. This challenged the systems with a much longer distance between the anaphors and their antecedents.

8.2 Task 2 – Bridging Anaphora

Two teams submitted their results to Task 2, with UTD_NLP participating in all three phases and DFKI-INRIA only participating in the antecedent selection (Gold A) setting. The entity F1 scores for each sub-corpora together with the macro-average of those scores, the latter was used for ranking the systems.

Two teams submitted a total of 102 runs to the leaderboard for three different settings (67 runs for Pred, 5 runs for Gold M and 30 runs for Gold A).

Team	LIGHT	AMI	PERS.	SWBD.	Avg.
Eval DD (Gold A)					
UTD_NLP	**52.40**	**72.50**	**69.61**	**72.11**	**66.66**
DFKI-INRIA	44.95	56.54	62.79	0.00	41.07
Baseline	40.07	39.89	51.43	37.72	42.28
Eval DD (Gold M)					
UTD_NLP	**38.38**	**55.12**	**54.89**	**49.83**	**49.56**
DFKI-INRIA	35.91	47.13	48.24	0.00	32.82
Baseline	18.14	22.95	30.15	21.37	23.15
Eval DD (Pred)					
UTD_NLP	**37.09**	**53.31**	**54.59**	**49.76**	**48.69**
DFKI-INRIA	36.82	50.09	47.04	0.00	33.49
Baseline	10.94	17.39	16.61	13.30	14.56

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

This makes bridging (Task 2) overtaking the identity resolution (Task 1) becomes the most popular task of this year's shared task in terms number of runs submitted to the leaderboard. Table 4 introduces the results of each phases. For the predicted mention setting (Pred), where the systems need to predict both the mentions and the bridging relations, the baseline only achieved a score of 5% on average. The task is very challenging given that only a limited amount of training data is available and the complexity of the bridging task itself. Yet the best result from UTD_NLP quadrupled the ones of the baseline. With the help of available gold mention (Gold M), both the baseline and the UTD_NLP performance further improved slightly by 3-5%. The small improvements achieved by using the gold mentions indicate that 1. the mentions predicted by the systems are not substantially different from the gold mentions; 2. the bridging task remains very challenging even though the gold mentions are provided. In the gold anaphor setting (Gold A) where the gold bridging anaphors are made available in addition to the gold mentions, the system performance increased dramatically. The baseline performance is more than tripled and the best results are 20% higher than the ones of the gold mention (Gold M) setting. Over the four sub-corpora, the PERSUASION seems to be the easiest corpus, both baseline and the participating systems achieved the best results on this corpus. The system results on the other three sub-corpus vary from system to system, in general, no clear distinction between them.

8.3 Task 3 – Discourse Deixis

For Task 3, two teams (UTD_NLP and DFKI-INRIA) participated in all three phases. In total, we received 72 runs from them, in which 30 runs were submitted to the predicted mention setting (Pred), 34 runs for the gold mention setting (Gold M) and 8 runs for the gold anaphor setting (Gold A). The UTD_NLP team submitted results for all four sub-corpora whereas the DFKI-INRIA team submitted predictions for three sub-corpora leaving the SWITCHBOARD behind. We report the CoNLL average F1 for each sub-corpora and rank the systems using the mean of those scores (see Table 5).

For the predicted mention setting, the baseline system achieved a score of around 15% for all four sub-corpora, both participating systems achieved much better results than the baseline. The performances are relatively close for LIGHT and AMI, and for PERSUASION, the UTD_NLP is 7% better than the DFKI-INRIA team. The best performing system achieved CoNLL average F1 scores on or above 50% for all sub-corpora evaluated, the only exception is the LIGHT which is more than 10% lower than other corpora. In the gold mention setting (Gold M), the baseline does improve largely (9%) by further filtering the heuristic anaphors with the gold mentions. However, the additionally available gold mentions do not improve largely the performance of the participating systems. The performance of the DFKI-INRIA team on LIGHT and AMI even dropped slightly. Finally, in the gold anaphor setting (Gold A), the naive baseline already achieved a score above 40%, and the best participating system achieved an F1 above 66% on average. This suggests the identification of discourse deixis anaphor remains challenging. Overall, all the systems outperform the baseline by a large margin in all the sub-corpora they participated.

8.4 Discussion

Since this is the second year of the shared task, we adopted many valuable assets from the first year, such as the scorer, the code to set up the CodaLab and the baselines etc. For this year, one of the main focus becomes to improve the quality of the annotation. We managed to release the revised version of the RST portion of the ARRAU 3 data that serves as the main training data for the shared task. In addition, we also annotated brand new test sets for all sub-corpora and revised the dev/test sets from last year to make them train/dev sets respectively. The consistency of the annotation has been largely improved for this year's shared task data and this makes the corpus of higher quality. We also managed to release most of the data as scheduled. Apart from the data, we also introduced the gold anaphor settings for bridging and discourse deixis tasks to allow the participants to develop systems focused on the antecedent selection subtask. To adapt to the new phase, we extended the baselines from last year to the gold settings.

In terms of the results, although the test sets are not the same as last year, the baseline performance remains similar is a good indication that the hardness of the tasks does not change much. In comparison with last year, we noticed some improvements for both bridging and discourse deixis tasks. The performance on the bridging task improved 3-5% on average and for discourse deixis, we saw large improvements of 6% and 10% for the gold/predicted mention settings respectively. Apart from more advanced systems being used, the additional in-domain training set available this year might also play a role in the improvements. By contrast, the best performances on identity resolution are similar to last year's. This might as a result of the development set that was already used for training by the best-performing system from last year. Hence the settings are not that different between the two years.

Finally, we would like to thank all participants for making a great effort to push further the performances on all the individual tasks. And congratulate them for outperforming the baselines by large margins.

9 Conclusion and Future Work

In this paper we presented a general overview of the CODI-CRAC 2022 shared task. Like the first shared task in this series, CODI-CRAC 2022 focused on resolving three types of anaphoric relations in dialogues: identity, bridging reference, and discourse deixis.

Based on the feedback from participants to the first task, in this second event we released the annotation guidelines beforehand so that participants could know exactly how the data had been annotated. In addition, we re-checked the data newly annotated for the first edition (now available for training and development, so that participants could

do some in-domain training as well), and using a larger group of annotators, which resulted in an hopefully more objective annotation. New test data in the four new dialogue domains was also annotated.

The participant systems outperformed the baselines on virtually all tasks and settings, although a clear difference in performance could be observed for bridging reference between pure resolution and resolution + identification. (Interestingly, we didn't observe much difference in performance between the 'Gold Mention' and 'Predicted' settings for either bridging nor discourse deixis.) A clear difference was observed between the results on the AMI datasets and on the other datasets for identity anaphora and bridging reference, possibly due to greater length of the documents in AMI.

Acknowledgments

We are very grateful to Maris Camilleri, Paloma Carretero Garcia, Taiqi He and Katherine Zhang not only for the annotation but for their extensive analysis of the annotation scheme and the data, resulting in useful discussions. We would also like to thank the Linguistic Data Consortium, who very generously made the ARRAU and SWITCHBOARD data available to the participants to the competition.

The work of Massimo Poesio, Juntao Yu, Maris Camilleri and Paloma Carretero Garcia was funded in part by the the DALI project, ERC Grant 695662; in part by the EPSRC project ARCIDUCA, grant number EP/W001632/1; and in part by HITS Heidelberg. The work of Lori Levin, Taiqi He and Katherine Zhang was funded by the Language Technologies Institute at Carnegie Mellon University.

References

Tatiana Anikina, Cennet Oguz, Natalia Skachkova, Siyu Tao, Sharmila Upadhyaya, and Ivana Kruijff-Korbayova. 2021. Anaphora resolution in dialogue: Description of the DFKI-TalkingRobots system for the CODI-CRAC 2021 shared-task. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 32–42, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

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

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.

Herbert H. Clark and Susan E. Brennan. 1990. Grounding in communication. In L. B. Resnick, J. Levine, and S. D. Behrend, editors, *Perspectives on Socially Shared Cognition*. APA.

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).

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.

Peter A. Heeman and Graeme Hirst. 1995. Collaborating on referring expressions. *Computational Linguistics*, 21(3):351–382.

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.

Sopan Khosla, Juntao Yu, Ramesh Manuvinakurike, Vincent Ng, Massimo Poesio, Michael Strube, and Carolyn Rosé. 2021. The CODI-CRAC 2021 shared task on anaphora, bridging, and discourse deixis in dialogue. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hideo Kobayashi, Shengjie Li, and Vincent Ng. 2021. Neural anaphora resolution in dialogue. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 16–31, Punta Cana, Dominican Republic. Association for 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, 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.

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.

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*.

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 Sadat Moosavi, and Massimo Poesio. 2022. Scoring coreference chains with split-antecedent anaphors.

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 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 Hannes Rieser. 2010. Completions, coordination, and alignment in dialogue. *Dialogue and Discourse*, 1(1):1–89.

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. 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 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.

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

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.