The CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis Resolution in Dialogue: A Cross-Team Analysis

Shengjie Li* and **Hideo Kobayashi*** and **Vincent Ng**
Human Language Technology Research Institute
University of Texas at Dallas
Richardson, TX 75083-0688
{sxl180006,hideo,vince}@hlt.utdallas.edu

Abstract

The CODI-CRAC 2021 shared task is the first shared task that focuses exclusively on anaphora resolution in dialogue and provides three tracks, namely entity coreference resolution, bridging resolution, and discourse deixis resolution. We perform a cross-task analysis of the systems that participated in the shared task in each of these tracks.

1 Introduction

The CODI-CRAC 2021 shared task (Khosla et al., 2021), which focuses on anaphora resolution in dialogue, provides three tracks, namely entity coreference resolution, bridging resolution, and discourse deixis/abstract anaphora resolution. Among these three tracks, bridging resolution and discourse deixis resolution are relatively under-studied problems. This is particularly so in the context of dialogue processing. This shared task is therefore of potential interest to researchers in the discourse and dialogue communities, particularly researchers in anaphora resolution who intend to work on problems beyond identity coreference.

Our goal in this paper is to perform a cross-team analysis of the systems participating in the three tracks of the shared task. Our analysis is partly quantitative, where we attempt to draw conclusions based on statistics computed using the outputs of the systems, and partly qualitative, where we discuss the strengths and weaknesses of the systems based on our manual inspection of these outputs. While several attempts have been made to perform an analysis of different coreference systems (e.g., Kummerfeld and Klein (2013), Lu and Ng (2020)), we note that conducting an insightful analysis of these systems is inherently challenging for at least two reasons. First, for entity coreference resolution and discourse deixis resolution, the latter of which is treated as a general case of event coreference, the system outputs on which we perform our analysis is in the form of clusters. Hence, we do not have information about which *links* were posited by a system and used to create a given cluster. This makes it impossible to pinpoint the mistakes (i.e., the erroneous linking decisions) made by a system and fundamentally limits our ability to *explain* the behavior of a system. Second, even if we could pinpoint the mistakes, existing models for anaphora resolution have become so complex that it is virtually impossible to explain why a particular mistake was made. For instance, a mention extraction component is so closely tied to a resolution model that it is not always possible to determine whether a mistake can be attributed to erroneous mention extraction or resolution. Worse still, since the participants have the freedom to partition the available training and development datasets in any way they want for model training and parameter tuning and are even allowed to exploit external training corpora, it makes it even harder to determine whether a system performs better because of a particular way of partitioning the data or because external training data are used.

The rest of this paper is structured as follows. The next three sections describe our cross-team analysis for the three tracks, namely entity coreference (Section 2), bridging (Section 3), and discourse deixis (Section 4). We present our conclusions and observations in Section 5.

2 Entity Coreference Resolution

In this section, we analyze the results of the four teams that participated in the anaphora resolution track and submitted a shared task paper, namely the team from Emory University (Xu and Choi, 2021) (henceforth Emory), the team from the University of Texas at Dallas (Kobayashi et al., 2021) (henceforth UTD), the team from Korea University (Kim et al., 2021) (henceforth KU), and the DFKI team (Anikina et al., 2021) (henceforth DFKI).

*Equal contribution

Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue, pages 71–95
Punta Cana, Dominican Republic, November 10, 2021.

	LIGHT			AMI			Persuasion			Switchboard		
	P	R	F	P	R	F	P	R	F	P	R	F
Emory	89.2	92.5	90.8	82.2	90.2	86.0	90.6	90.7	90.6	85.3	89.8	87.5
UTD	92.3	91.6	92.0	86.6	78.6	82.4	91.3	89.7	90.5	89.2	86.1	87.6
KU	85.6	92.8	89.1	79.4	89.3	84.0	83.3	92.5	87.7	78.7	89.8	83.8
DFKI	84.8	82.6	83.7	75.4	65.8	70.3	79.8	77.5	78.6	79.3	77.9	78.6

Table 1: Entity coreference resolution: mention extraction results.

2.1 Mention Extraction

Since mention extraction has a large impact on coreference resolution performance (Pradhan et al., 2011, 2012), let us first consider the mention extraction performance of the participating systems.

Table 1 presents the mention extraction results in the standard manner, expressing the results of each system on each corpus in terms of recall, precision, and F-score. Specifically, a mention is considered correctly detected if it has an exact match with a gold mention in terms of boundary. In terms of F-score, Emory and UTD achieve comparable performance, and both of them outperform KU and DFKI. Except for DFKI, all the systems achieve an average mention extraction F-score of more than 85%. These mention extraction results are much better than those achieved by traditional coreference resolvers, and are consistent with Lu and Ng's (2020) observation that mention detection performance has improved significantly over the years, particularly after the introduction of span-based neural coreference models (Lee et al., 2017, 2018). Considering the recall and precision numbers, we see that Emory and KU are recall-oriented, whereas UTD and DFKI are precision-oriented. Specifically, Emory and KU achieve the highest recall, whereas UTD achieves the highest precision.

To gain additional insights into the mention extraction results achieved by these systems, we present these results from a different point of view in Table 2. We first divide the mentions into 10 groups, which are shown in Table 3. As can be seen, the first nine groups focus on different kinds of pronouns, whereas the last group is composed of non-pronominal mentions. We note that the classification of the mentions in the test corpus into these 10 groups is not error-free: since we rely on part-of-speech tags and the surface forms to identify pronouns, words that appear in the corpus such as "well" (which corresponds to "we'll") and "were" (which corresponds to "we're") should belong to Group 1 but are being misclassified as non-pronominal.

Consider first Table 2a, where results are aggregated over the four datasets. The "%" and "count" columns show the percentage and number of gold mentions that belong to each group. "none" shows the fraction of mentions that are not detected by any of the four participating systems. E (Emory), T (UTD), K (KU), and D (DFKI) show the percentage of gold mentions successfully extracted by each of these systems. E-only, T-only, K-only, and D-only show the percentage of gold mentions that are successfully extracted by exactly one of the systems. For instance, E-only shows the fraction of mentions successfully extracted by the Emory resolver but not the other three.

A few points deserve mention. First, despite the fact that Group 10 (non-pronominal mentions) is the largest group, approximately half of the mentions are pronominal. The largest pronoun groups are Group 1 (1st and 2nd person pronouns (e.g., "I", "we")), Group 3 (3rd person ungendered pronouns (e.g., "it", "they")), Group 5 (reflexive pronouns (e.g., "myself", "yourself")), and Group 7 (demonstrative pronouns (e.g., "this", "that")). This should not be surprising given the prevalence of these pronouns in dialogue, but their prevalence suggests the importance of pronoun resolution in the shared task. Second, considering the "only" columns, we see that the percentage of mentions that are uniquely identified by one of the systems is relatively small. Third, Emory and KU extract more gold mentions than UTD and DFKI. As can be seen in the "Overall" row, Emory and KU manage to extract more than 90% of the gold mentions. These results are consistent with those shown in Table 1. Perhaps the biggest difference among the systems lies in the extraction of non-pronominal mentions: Group 10 is the group for which Emory and KU clearly demonstrate superior extraction performance.

While Table 2a focuses on gold mention extraction, Table 2b focuses on the extraction of erroneous mentions. Specifically, we take the union of the set of erroneous mentions extracted by the

(a) Coverage of gold mentions

Group	%	count	none	E	T	K	D	E-only	T-only	K-only	D-only
Overall	100.0	19051	4.1	90.5	86.1	90.7	76.0	1.0	0.5	1.5	0.9
1	28.5	5433	0.0	99.8	99.6	99.0	99.7	0.0	0.0	0.0	0.0
2	1.6	308	0.0	99.4	99.7	99.4	100.0	0.0	0.0	0.0	0.3
3	8.8	1670	0.0	98.3	97.1	99.8	100.0	0.0	0.0	0.0	0.1
4	5.2	986	0.0	98.5	98.7	99.4	94.7	0.0	0.0	0.5	0.0
5	0.3	65	0.0	95.4	89.2	93.8	100.0	0.0	0.0	0.0	0.0
6	0.9	172	2.3	79.7	68.6	88.4	56.4	0.6	0.0	3.5	3.5
7	5.1	974	1.3	97.8	96.6	96.9	0.0	0.2	0.1	0.5	0.0
8	0.3	61	6.6	70.5	63.9	78.7	45.9	0.0	0.0	0.0	3.3
9	0.6	115	3.5	77.4	75.7	94.8	93.0	0.0	0.9	0.9	0.0
10	48.6	9267	8.1	82.3	74.0	82.4	63.3	2.0	0.9	2.9	1.7

(b) Coverage of wrong mentions

Group	%	count	E	T	K	D	E-only	T-only	K-only	D-only
Overall	100.0	7072	39.0	26.7	57.7	52.4	10.0	4.2	18.8	20.7
1	0.5	32	75.0	71.9	87.5	87.5	3.1	0.0	9.4	0.0
2	0.0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	6.2	439	58.3	45.3	99.8	100.0	0.0	0.0	0.0	0.2
4	0.9	67	61.2	58.2	92.5	55.2	0.0	0.0	17.9	3.0
5	0.0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	2.4	173	23.7	16.8	66.5	69.4	3.5	0.6	22.0	28.9
7	0.7	52	82.7	53.8	50.0	0.0	32.7	3.8	5.8	0.0
8	5.3	376	17.0	14.1	39.4	90.7	0.5	0.5	5.1	56.6
9	3.7	262	16.4	9.5	99.2	83.2	0.4	0.0	8.8	0.0
10	80.2	5671	39.6	26.3	52.9	44.5	12.0	5.1	21.7	21.1

Table 2: Entity coreference resolution: per-class mention extraction results.

Group	**Description**
1	1st and 2nd person pronouns
2	3rd person gendered pronouns
3	3rd person ungendered pronouns
4	Possessive pronouns
5	Reflexive pronouns
6	Indefinite pronouns
7	Demonstrative pronouns
8	Relative and interrogative pronouns
9	Other pronouns
10	Non-pronominal noun phrases

Table 3: Division of mentions into groups.

four resolvers and compute statistics based on the resulting set, which we refer to as S. The columns in Table 2b can be interpreted in the same way as those in Table 2a. For instance, E, T, K, and D show the percentage of mentions in S extracted by each of the systems, and E-only, T-only, K-only, and D-only show the percentage of mentions in S extracted by exactly one of the systems.

A few points deserve mention. First, approximately 80% of the erroneous mentions belong to Group 10. This should perhaps not be surprising given that the extraction of non-pronominal mentions, which are often composed of multiple tokens, is typically more challenging than that of pronouns. Note that a complication involved in the extraction process concerns the detection of non-referring mentions: according to the shared task guidelines, any non-referring mention extracted will be considered erroneous. The second largest group is Group 3, whose mentions account for 6.2% of the number of erroneous mentions. This again should not be surprising. This group is composed of pronouns such as "it", many of which may not be referring because of its use as an expletive or a pleonastic pronoun. Second, considering the "Overall" row, we can see that UTD has the smallest coverage of erroneous mentions, which translates to a higher mention extraction precision., whereas KU has the highest coverage of erroneous mentions. While Table 2a shows that Emory and KU both achieve high mention extraction recall, Table 2b shows that KU does so at the expense of precision and that Emory is clearly superior to KU for mention extraction. While DFKI's coverage of gold mentions is the lowest among the four systems, its coverage of erroneous mentions is relatively high. Third, considering the group-specific results, we can get a better idea of what makes a particular system better for mention extraction. UTD extracts fewer erroneous mentions

	MUC			B^3			$CEAF_e$			
	P	R	F	P	R	F	P	R	F	CoNLL
					LIGHT					
Emory	90.6	89.5	90.0	74.2	84.9	79.2	68.1	75.9	71.8	80.3
UTD	87.8	89.1	88.5	72.7	82.6	77.3	74.5	73.9	74.2	79.6
KU	89.3	75.3	81.7	64.9	54.9	59.5	55.7	78.4	65.1	68.8
DFKI	86.2	70.9	77.8	75.3	59.0	66.2	52.8	64.1	57.9	67.3
					AMI					
Emory	72.4	69.0	70.7	57.2	66.2	61.4	53.5	67.9	59.8	64.0
UTD	66.7	65.5	66.0	48.5	58.3	53.0	51.1	46.4	48.6	57.4
KU	69.0	53.0	60.0	63.6	49.5	55.7	46.5	72.4	56.6	57.4
DFKI	59.1	41.9	49.1	48.1	41.3	44.4	37.0	39.7	38.3	43.9
					Persuasion					
Emory	80.4	88.0	84.0	76.9	82.5	79.6	74.7	68.8	71.6	78.4
UTD	78.7	87.8	83.0	76.6	80.4	78.5	76.2	74.8	75.5	77.5
KU	76.5	77.2	76.9	65.6	70.4	67.9	61.9	73.1	67.0	70.6
DFKI	69.8	65.2	67.4	65.2	55.1	59.7	52.6	52.7	52.6	59.9
					Switchboard					
Emory	82.2	79.1	80.6	72.4	76.5	74.4	64.1	73.4	68.5	74.5
UTD	77.5	79.5	78.5	70.7	74.3	72.4	71.5	69.0	70.2	72.6
KU	80.9	66.4	72.9	66.3	58.2	62.0	52.0	75.8	61.7	65.5
DFKI	62.7	60.2	61.4	55.6	56.9	56.3	42.8	43.1	43.0	53.5

Table 4: Entity coreference resolution: official resolution results.

than other teams except for Groups 4 and 7, both of which are relatively small. By contrast, KU extracts considerably more erroneous possessive pronouns (Group 4) than other teams, Emory extracts considerably more erroneous demonstrative pronouns (Group 7) than other teams, DFKI extracts more erroneous relative and interrogative pronouns (Group 8) than other teams, and both KU and DFKI extract considerably more erroneous indefinite pronouns (Group 6) and other pronouns (Group 9) than other teams. Finally, considering the "only" columns, we see that 10% of the erroneous mentions are only extracted by Emory, 18.8% by KU, and 20.7% by DFKI. This shows that the systems are quite different in terms of mention extraction.

2.2 Resolution

Next, we consider the coreference results. Table 4 shows the official results obtained using the official scorer. These results are expressed in terms of MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), and $CEAF_e$ (Luo, 2005) recall (R), precision (P), and F-score (F), as well as the CoNLL score, which is the unweighted average of the MUC, B^3, and $CEAF_e$ F-scores.

The four participating systems show a clear difference in performance in terms of CoNLL F-score: Emory performs the best, UTD and KU rank second and third respectively, and DFKI achieves the lowest performance. The performance difference between Emory and UTD is smaller compared to that between any other pair of systems: UTD underperforms Emory roughly by 0.7–6.6% in the CoNLL score. This could be explained in part by the fact that both systems were built upon coref-hoi, which is Xu and Choi's (2020) entity coreference system. Nevertheless, as we will see below, the two systems behave quite differently.

Next, we consider the performance of these systems w.r.t. each scorer. As a link-based metric, MUC focuses on link identification and does not reward successful identification of singleton clusters. Hence, by looking at the MUC recall, we can get a better idea of how well a resolver does in terms of link identification. As can be seen, Emory and UTD are substantially better than KU and DFKI in terms of link identification. To gain a better understanding of the extent to which the singleton clusters and the non-singleton clusters contribute to the overall performance of the systems, we show the corresponding results in Table 5. Specifically, in the "F" columns we show the CoNLL score. The "ns-F" columns show the CoNLL scores obtained by removing the singleton clusters from the output prior to scoring, meaning that the scorers are applied to score only the non-singleton clusters. Similarly, the "s-F" columns show the CoNLL scores obtained by removing the non-singleton clusters from the output prior to scoring, effectively allow-

	LIGHT			AMI			Persuasion			Switchboard		
	F	ns. F	s. F	F	ns. F	s. F	F	ns. F	s. F	F	ns. F	s. F
Emory	80.3	60.5	32.4	64.0	45.0	29.3	78.4	56.6	35.8	74.5	53.6	33.7
UTD	79.6	60.2	32.9	57.4	42.9	25.2	77.5	56.5	35.3	72.6	53.7	32.5
KU	68.8	49.4	28.0	57.4	38.5	27.5	70.6	48.0	33.1	65.5	44.8	30.7
DFKI	67.3	51.2	27.0	43.9	32.2	19.1	59.9	43.2	27.3	53.5	41.6	20.7

Table 5: Entity coreference resolution: results on singleton and non-singleton cluster identification.

ing the scorers to score only the singleton clusters.

As we can see from the results in Table 5, Emory and UTD achieve comparable performance w.r.t. both non-singleton and singleton cluster identification, except on the AMI dataset where Emory clearly demonstrates its superior performance w.r.t. both tasks. In addition, while Emory and UTD are generally better than KU w.r.t. both tasks, the difference stems more from non-singleton cluster identification than singleton cluster identification. Comparing KU and DFKI, we see that KU is better than DFKI on both tasks on all but the LIGHT dataset.

Next, we measure system performance, specifically link identification performance, at the *pairwise* level. For each non-singleton coreference cluster in the gold output, we extract every pair of mentions in the cluster, and denote the set of pairs extracted from all non-singleton clusters as G_p. We similarly extract all the pairs from the non-singleton clusters produced by each of the four systems, and denote the resulting sets as E_p, T_p, K_p and D_p. The recall, precision, and F-scores in Table 6 are computed based on the pairwise links in these sets.

From the "Overall" row in Table 6, we can see that approximately 120K pairs can be extracted from the gold clusters of the four test datasets. As we can see, except for KU, all systems have higher recall than precision. In particular, UTD has the highest recall but the lowest precision. Comparing Emory and UTD, we see that while the two systems achieve comparable recall, UTD's precision is much lower, which in turn results in a lower F-score. Though precision-oriented, KU's precision is not as high as Emory's, which has the highest precision among the systems.

To gain a better understanding of how these systems perform w.r.t. identifying difficult vs. easy links, we divide the pairs into three groups based on an intuitive notion of hardness. Group A is composed of pairs where the two mentions are lexically identical. A pair appears in Group B if (1) both mentions in the pair are pronouns or (2) both mentions are non-pronominal and have a content word overlap. Finally, a pair appears in Group C if (1) the anaphor is pronominal but the antecedent is not or (2) the two mentions have no content word overlap.

Results are shown in the rows labeled A, B, and C. The easiest links (Group A) account for nearly half of all pairs while the hardest links (Group C) have a much lower representation, accounting for only 15% of all pairs. Emory achieves the best results in all three groups, indicating its robustness in identifying both easy and hard links. UTD ranks second in Groups A and B, and the performance gap between Emory and UTD widens as hardness increases. DFKI ranks third in Groups A and B, and largely fails to identify the links in Group C. Finally, while KU does poorly for Groups A and B, it performs slightly better than UTD w.r.t. Group C. We speculate that KU chooses to resolve only those pairs it is most confident about regardless of hardness, yielding a low recall but a higher precision.

To understand how well each system does in resolving the anaphors in each of the 10 groups we defined earlier, we show the per-group results in the rows labeled 1 through 10. As can be seen, the links involving 1st or 2nd pronouns as anaphors (Group 1) form the largest group, accounting for 70% of all links. This is followed by links involving non-pronominal mentions (Group 10) and possessive pronouns (Group 4), both of which account for slightly more than 10% of all links. Emory achieves the best performance on these three largest groups. Comparing Emory and UTD, we can see that the two systems are indeed different: Emory achieves a higher precision than UTD on all 10 groups, and its precision and recall are both higher than UTD's on Groups 6, 7, and 10. Comparing KU and DFKI, we see that DFKI outperforms KU in resolving 1st and 2nd pronouns (Group 1), possessive pronouns (Group 4) and reflexive pronouns (Group 5). Though achieving the lowest overall performance, KU outperforms UTD in resolving

Group	%	count	Emory			UTD			KU			DFKI		
			P	R	F	P	R	F	P	R	F	P	R	F
Overall	100.0	120045	60.4	78.9	68.4	45.2	80.2	57.8	57.9	30.6	40.1	49.1	58.7	53.5
A	46.4	55702	63.4	84.3	72.4	55.4	90.5	68.7	64.6	28.1	39.2	54.4	79.9	64.8
B	38.6	46357	62.8	80.7	70.6	41.3	83.1	55.2	55.5	35.5	43.3	44.5	54.9	49.1
C	15.0	17986	44.5	57.2	50.1	25.8	41.0	31.6	48.4	25.9	33.7	10.1	2.5	4.0
1	70.0	84000	62.2	84.3	71.6	48.1	87.0	61.9	57.5	28.2	37.8	49.1	71.1	58.1
2	2.4	2835	81.6	68.8	74.7	59.4	78.6	67.6	87.1	51.2	64.5	58.4	35.1	43.9
3	5.0	5990	47.2	53.9	50.3	27.1	59.1	37.2	57.1	44.2	49.8	29.9	14.0	19.1
4	10.1	12124	62.3	82.2	70.9	46.1	84.6	59.7	60.5	42.2	49.7	55.3	62.6	58.7
5	0.7	862	67.2	68.6	67.9	56.9	74.9	64.7	49.3	20.6	29.1	51.1	49.9	50.5
6	0.1	158	38.4	27.2	31.9	24.1	21.5	22.7	59.0	22.8	32.9	7.7	0.6	1.2
7	0.9	1051	37.9	40.9	39.3	18.7	34.1	24.1	44.8	29.6	35.6	nan	0.0	nan
8	0.0	37	15.6	13.5	14.5	11.5	18.9	14.3	23.3	18.9	20.9	0.0	0.0	nan
9	0.2	259	50.1	74.9	60.1	28.2	78.4	41.5	37.5	47.5	41.9	26.7	3.1	5.5
10	10.6	12729	49.2	58.7	53.5	33.1	46.5	38.6	52.9	25.2	34.2	31.3	6.6	10.9
1A	39.7	47641	63.0	86.7	73.0	56.0	93.2	69.9	62.1	25.1	35.7	53.8	88.1	66.8
1B	26.2	31414	64.2	84.2	72.8	42.0	85.2	56.2	54.4	33.7	41.6	41.9	56.4	48.1
1C	4.1	4945	43.4	61.6	50.9	21.7	39.2	27.9	46.1	23.6	31.2	2.0	0.5	0.8
2A	0.9	1108	91.3	74.2	81.9	66.4	94.7	78.0	94.2	58.8	72.4	79.2	46.4	58.5
2B	0.9	1097	79.6	64.4	71.2	54.2	85.9	66.5	85.2	49.2	62.4	46.4	32.7	38.4
2C	0.5	630	70.1	67.0	68.5	54.6	37.5	44.4	76.1	41.4	53.6	43.5	19.5	26.9
3A	1.9	2322	53.4	63.3	57.9	36.3	67.4	47.2	67.0	53.7	59.6	56.0	22.3	31.9
3B	1.1	1289	40.7	45.5	43.0	13.8	51.0	21.8	46.3	40.6	43.2	22.3	9.3	13.1
3C	2.0	2379	44.2	49.3	46.6	33.0	55.3	41.4	53.4	36.9	43.7	14.9	8.4	10.8
4A	1.1	1271	71.3	87.3	78.5	55.9	91.3	69.3	68.9	46.7	55.7	68.4	75.0	71.5
4B	7.9	9499	63.4	83.9	72.2	46.9	88.3	61.2	60.9	43.0	50.4	54.9	69.3	61.3
4C	1.1	1354	47.2	65.4	54.8	31.1	52.1	39.0	49.7	33.1	39.7	15.7	4.2	6.6
5A	0.0	11	42.9	27.3	33.3	80.0	36.4	50.0	0.0	0.0	nan	46.2	54.5	50.0
5B	0.6	750	68.4	73.1	70.7	57.5	79.2	66.6	50.0	20.5	29.1	52.8	56.3	54.5
5C	0.1	101	56.3	39.6	46.5	49.5	47.5	48.5	48.0	23.8	31.8	6.9	2.0	3.1
6A	0.1	69	50.0	37.7	43.0	32.1	37.7	34.7	69.7	33.3	45.1	8.3	1.4	2.5
6B	0.0	38	0.0	0.0	nan	0.0	0.0	nan	0.0	0.0	nan	0.0	0.0	nan
6C	0.0	51	39.5	33.3	36.2	17.8	15.7	16.7	59.1	25.5	35.6	nan	0.0	nan
7A	0.2	280	53.8	60.0	56.8	38.6	51.1	44.0	59.8	53.6	56.5	nan	0.0	nan
7B	0.3	318	33.0	36.2	34.5	12.5	36.8	18.6	34.8	21.7	26.7	nan	0.0	nan
7C	0.4	453	30.9	32.5	31.7	16.1	21.6	18.5	37.6	20.3	26.4	nan	0.0	nan
8A	0.0	1	0.0	0.0	nan	0.0	0.0	nan	100.0	100.0	100.0	0.0	0.0	nan
8B	0.0	6	14.3	16.7	15.4	5.6	16.7	8.3	16.7	16.7	16.7	0.0	0.0	nan
8C	0.0	30	16.7	13.3	14.8	14.6	20.0	16.9	21.7	16.7	18.9	nan	0.0	nan
9A	0.0	42	49.2	76.2	59.8	42.7	76.2	54.7	40.0	33.3	36.4	46.7	16.7	24.6
9B	0.1	125	62.6	87.2	72.9	30.4	92.0	45.7	38.2	58.4	46.2	0.0	0.0	nan
9C	0.1	92	35.8	57.6	44.2	21.0	60.9	31.2	35.3	39.1	37.1	100.0	1.1	2.2
10A	2.5	2957	70.3	70.0	70.2	64.9	67.3	66.1	81.5	34.4	48.4	75.3	18.5	29.7
10B	1.5	1821	43.8	49.9	46.6	33.7	52.6	41.0	43.5	25.5	32.1	30.0	13.8	18.9
10C	6.6	7951	44.1	56.6	49.6	24.8	37.3	29.8	46.0	21.8	29.6	3.7	0.5	0.9

Table 6: Entity coreference resolution: resolution results at the pairwise level.

3rd person ungendered pronouns (Group 3), indefinite pronouns (Group 6), demonstrative pronouns (Group 7), relative and interrogative pronouns (Group 8), and other pronouns (Group 9). The remaining rows of the table show the results when each of the ten groups is further subdivided into three groups based on the three levels of hardness. Space limitations preclude a discussion of these results, however.

Table 7a shows each system's coverage of gold coreferent pairs. The rows can be interpreted in the same way as those in Table 6, whereas the columns can be interpreted in the same way as those in Table 2. As a quick reminder, "none" shows the percentage of gold pairs that are not extracted by any of the four systems; "E", "T", "K", and "D" show the percentage of gold pairs extracted by each of the four systems; and the "x-only" columns show the percentage of gold pairs that are extracted only by system x.

A few points deserve mention. First, consider the "none" results. As can be seen, only 10.5% of the links are not recovered by any of the four systems. Taking into account link hardness, we see that 31.9% of the hardest links (Group C) are not extracted while only 4.9% of the easiest links (Group A) are not extracted. These results provide suggestive evidence that our intuition notion of link hardness is consistent with what a resolver would perceive as hard. Considering the per-group results (Group i, where $1 \leq i \leq 10$), approximately 65% of the links involving indefinite pronouns (Group 6) and relative and interrogative pronouns (Group 8), 47% of the links involving demonstrative pronouns (Group 7), and 29% of the links involving 3rd person ungendered pronouns (Group 3) and non-pronominal mentions (Group 10) are missed by all four systems. These are traditionally the harder groups of anaphors to resolve. Second, consider the "x-only" results. While the "Overall" results suggest that Emory and UTD may not very different from each other in terms of the links they recover, the "x-only" results suggest otherwise. Specifically, 18% of the hardest links (Group C) and 16.6% of the non-pronominal links (Group 10) are uniquely identified by Emory, whereas 15.1% of the links involving 3rd person gendered pronouns (Group 2) are uniquely identified by UTD.

While Table 7a focuses on the extraction of gold coreferent pairs, Table 7b focuses on pairs that are erroneously posited as coreferent. Specifically, we take the union of the set of pairs that are erroneously posited as coreferent by the four resolvers and compute statistics based on the resulting set, which we refer to as S. The columns in Table 7b can be interpreted in the same way as those in Table 2. For instance, E, T, K, and D show the percentage of pairs in S extracted by each of the systems, and the "x-only" columns show the percentage of pairs in S that are extracted only by system x.

As can be seen in Table 7b, approximately 19.4K erroneous links are established by the four systems, of which 60.2% are established by UTD, 37.7% by DFKI, 32% by Emory, and 13.8% by KU; in addition, 32.4% of these erroneous links are only established by UTD and 20.4% are only established by DFKI. Combining the results in Tables 7a and 7b, we see that UTD is the most aggressive among the four systems in link identification: it has the highest recall but the lowest precision, which is consistent with the results in Table 6, and a large percentage of erroneous links are only established by UTD. In fact, a closer examination of the results reveals that the percentage of erroneous links established by UTD is higher than that by any other system w.r.t. each of the ten groups and each of the three hardness groups. Furthermore, recall from Table 7a that Emory manages to correctly extract many hard links (Group C) and non-pronominal links (Group 10), but from Table 7b we can see that this success comes at the expense of extracting a fairly large number of erroneous links in these groups. Finally, the "only" columns show that the errors made by the four systems are quite different from each other.

2.3 Discussion

In this subsection, we manually analyze the pairwise links that are correctly and incorrectly established by the participating systems.

We begin by examining the coreference links that are not identified by any system. The mention pairs that are most frequently missed are: ("I", "I"), ("I", "you"), ("you", "I"), and ("it", it"), where the first mention in each pair is the anaphor and the second mention is its antecedent. It should perhaps not be surprising that these pairs all involve links between pronouns given their prevalence in spoken dialogues. More than 3000 instances of these four mention pairs are not extracted by any of the participating systems. The other major types

(a) Coverage of correct links

Group	%	count	none	E	T	K	D	E-only	T-only	K-only	D-only
Overall	100.0	120045	10.5	78.9	80.2	30.6	58.7	4.6	3.7	1.1	0.7
A	46.4	55702	4.8	84.3	90.5	28.1	79.9	1.3	2.6	0.6	0.6
B	38.6	46357	9.0	80.7	83.1	35.5	54.9	3.3	3.7	1.3	1.0
C	15.0	17986	31.9	57.2	41.0	25.9	2.5	18.0	7.1	2.4	0.1
1	70.0	84000	6.2	84.3	87.0	28.2	71.1	2.9	2.4	0.7	0.7
2	2.4	2835	9.7	68.8	78.6	51.2	35.1	5.4	15.1	0.5	1.1
3	5.0	5990	29.1	53.9	59.1	44.2	14.0	4.0	9.3	3.4	0.6
4	10.1	12124	7.1	82.2	84.6	42.2	62.6	3.2	2.7	1.3	1.1
5	0.7	862	13.6	68.6	74.9	20.6	49.9	2.6	5.3	1.6	2.7
6	0.1	158	65.8	27.2	21.5	22.8	0.6	4.4	1.9	1.9	0.0
7	0.9	1051	47.4	40.9	34.1	29.6	0.0	9.3	5.2	4.5	0.0
8	0.0	37	64.9	13.5	18.9	18.9	0.0	8.1	5.4	8.1	0.0
9	0.2	259	12.7	74.9	78.4	47.5	3.1	5.8	9.3	0.8	0.0
10	10.6	12729	29.3	58.7	46.5	25.2	6.6	16.6	7.8	2.4	0.2
1A	39.7	47641	3.1	86.7	93.2	25.1	88.1	0.6	1.6	0.3	0.7
1B	26.2	31414	7.3	84.2	85.2	33.7	56.4	3.4	3.0	1.0	0.8
1C	4.1	4945	29.4	61.6	39.2	23.6	0.5	22.3	6.1	1.9	0.0
2A	0.9	1108	1.8	74.2	94.7	58.8	46.4	1.7	17.6	0.0	1.0
2B	0.9	1097	8.3	64.4	85.9	49.2	32.7	1.9	20.2	0.3	1.5
2C	0.5	630	26.0	67.0	37.5	41.4	19.5	17.8	1.6	1.9	0.5
3A	1.9	2322	20.6	63.3	67.4	53.7	22.3	3.7	7.8	3.1	0.4
3B	1.1	1289	36.5	45.5	51.0	40.6	9.3	4.1	8.7	5.0	0.9
3C	2.0	2379	33.3	49.3	55.3	36.9	8.4	4.2	11.1	2.9	0.7
4A	1.1	1271	2.8	87.3	91.3	46.7	75.0	0.8	4.0	0.6	0.5
4B	7.9	9499	5.3	83.9	88.3	43.0	69.3	1.8	1.8	1.4	1.3
4C	1.1	1354	23.7	65.4	52.1	33.1	4.2	15.1	7.5	2.0	0.1
5A	0.0	11	36.4	27.3	36.4	0.0	54.5	9.1	0.0	0.0	18.2
5B	0.6	750	10.9	73.1	79.2	20.5	56.3	1.7	2.9	1.3	2.7
5C	0.1	101	30.7	39.6	47.5	23.8	2.0	7.9	23.8	4.0	1.0
6A	0.1	69	47.8	37.7	37.7	33.3	1.4	5.8	4.3	2.9	0.0
6B	0.0	38	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6C	0.0	51	64.7	33.3	15.7	25.5	0.0	5.9	0.0	2.0	0.0
7A	0.2	280	29.6	60.0	51.1	53.6	0.0	6.4	3.2	5.7	0.0
7B	0.3	318	47.8	36.2	36.8	21.7	0.0	8.5	7.2	5.3	0.0
7C	0.4	453	58.1	32.5	21.6	20.3	0.0	11.7	5.1	3.1	0.0
8A	0.0	1	0.0	0.0	0.0	100.0	0.0	0.0	0.0	100.0	0.0
8B	0.0	6	66.7	16.7	16.7	16.7	0.0	16.7	0.0	0.0	0.0
8C	0.0	30	66.7	13.3	20.0	16.7	0.0	6.7	6.7	6.7	0.0
9A	0.0	42	11.9	76.2	76.2	33.3	16.7	11.9	9.5	0.0	0.0
9B	0.1	125	4.0	87.2	92.0	58.4	0.0	4.0	8.8	0.0	0.0
9C	0.1	92	25.0	57.6	60.9	39.1	1.1	5.4	9.8	2.2	0.0
10A	2.5	2957	18.1	70.0	67.3	34.4	18.5	9.1	7.7	1.6	0.2
10B	1.5	1821	30.2	49.9	52.6	25.5	13.8	10.5	12.7	2.7	0.9
10C	6.6	7951	33.3	56.6	37.3	21.8	0.5	20.8	6.7	2.6	0.0

(b) Coverage of wrong links

Group	%	count	E	T	K	D	E-only	T-only	K-only	D-only
Overall	100.0	193827	32.0	60.2	13.8	37.7	9.6	32.4	7.3	20.4
A	34.6	66992	40.5	60.5	12.8	55.6	6.7	18.5	7.1	23.2
B	46.8	90794	24.4	60.4	14.6	35.0	7.1	37.7	7.4	22.3
C	18.6	36041	35.6	59.0	13.8	11.1	21.2	44.7	7.3	10.1
1	70.0	135593	31.7	58.2	12.9	45.6	7.6	27.6	7.3	24.4
2	1.1	2207	19.9	69.0	9.8	32.2	4.5	48.7	2.5	20.5
3	7.3	14074	25.7	67.5	14.1	14.0	12.9	53.9	5.7	11.6
4	9.4	18316	33.0	65.4	18.3	33.5	8.3	34.2	8.5	14.0
5	0.6	1069	26.9	45.7	17.1	38.5	8.8	27.6	8.2	33.0
6	0.1	168	41.1	63.7	14.9	7.1	20.8	46.4	5.4	6.0
7	1.2	2290	30.8	68.0	16.7	0.0	18.9	58.1	9.7	0.0
8	0.0	88	30.7	61.4	26.1	6.8	13.6	48.9	9.1	6.8
9	0.3	631	30.6	81.9	32.5	3.5	8.9	47.9	5.9	2.2
10	10.0	19391	39.9	61.7	14.8	9.5	21.3	42.9	7.4	7.3
1A	30.6	59401	40.8	58.9	12.3	60.8	5.9	15.2	7.4	25.5
1B	33.3	64556	22.9	57.3	13.7	38.0	6.7	35.6	7.4	25.9
1C	6.0	11636	34.2	60.2	11.8	11.0	21.2	46.7	6.2	10.3
2A	0.3	566	13.8	94.0	7.1	23.9	0.5	64.8	0.0	4.8
2B	0.6	1179	15.4	67.4	8.0	35.2	2.5	48.5	2.4	27.1
2C	0.2	462	39.0	42.4	17.7	34.6	14.5	29.2	5.8	22.9
3A	2.0	3866	33.1	71.0	15.9	10.5	14.7	51.4	4.8	6.5
3B	2.7	5227	16.4	78.2	11.6	8.0	8.6	69.2	5.3	6.5
3C	2.6	4981	29.7	53.6	15.4	23.0	16.1	39.9	6.9	20.9
4A	0.6	1197	37.3	76.6	22.4	36.8	4.9	33.2	7.4	6.5
4B	7.4	14350	32.0	66.3	18.3	37.6	5.9	32.5	8.5	15.6
4C	1.4	2769	35.8	56.4	16.4	11.0	22.1	43.2	8.7	9.3
5A	0.0	12	33.3	8.3	25.0	58.3	16.7	0.0	16.7	41.7
5B	0.5	942	26.9	46.6	16.3	40.1	7.5	27.3	7.5	34.3
5C	0.1	115	27.0	42.6	22.6	23.5	18.3	33.0	13.0	21.7
6A	0.0	74	35.1	74.3	13.5	14.9	8.1	48.6	1.4	12.2
6B	0.0	29	58.6	51.7	20.7	3.4	31.0	31.0	6.9	3.4
6C	0.0	65	40.0	56.9	13.8	0.0	30.8	50.8	9.2	0.0
7A	0.2	375	38.4	60.5	26.9	0.0	20.5	45.3	12.5	0.0
7B	0.6	1080	21.6	75.9	11.9	0.0	13.9	69.4	8.5	0.0
7C	0.4	835	39.3	61.1	18.3	0.0	24.6	49.2	10.1	0.0
8A	0.0	6	16.7	33.3	0.0	50.0	16.7	33.3	0.0	50.0
8B	0.0	26	23.1	65.4	19.2	11.5	7.7	57.7	3.8	11.5
8C	0.0	56	35.7	62.5	32.1	0.0	16.1	46.4	12.5	0.0
9A	0.0	55	60.0	78.2	38.2	14.5	14.5	30.9	5.5	0.0
9B	0.2	300	21.7	87.7	39.3	4.7	4.0	50.0	2.7	4.7
9C	0.1	276	34.4	76.4	23.9	0.0	13.0	48.9	9.4	0.0
10A	0.7	1440	60.7	74.7	16.0	12.4	15.4	27.6	1.8	5.4
10B	1.6	3105	37.5	60.7	19.4	19.0	15.9	39.1	8.9	10.5
10C	7.7	14846	38.4	60.7	13.7	7.3	23.0	45.1	7.7	6.9

Table 7: Entity coreference resolution: coverage of wrong links at the pairwise level.

of missing links involve demonstrative pronouns, wh-pronouns, and one-anaphora.

As for missing links that involve non-pronominal expressions, some appear to be easy to identify as the two mentions involved are synonyms, such as ("the super market", "the grocery store"), ("the school", "the college"), and ("kids", "children"). Since we do not examine the context in which they appear or consider how far apart they are from each other, we cannot conclude why these seemingly simple cases are being missed by all systems.

Some missing links are difficult to identify because the two mentions involved appear to have different semantic types. Examples include: ("your destination", "the king of the goolehops"), ("your donation", "the organization"), ("this very reputable charity", "you"), ("this school your son goes to", "private"), ("the battery", "the standard"), ("sixty-six", "the street"), and ("Michael", "a").

There are links that are missed because one or both of the mentions involved are simply not extracted. Examples include: ("your husband", "a wonderful man, you know, who treats me very, you know, with a, with as") and ("your grandfathers of past", "the kings of old: my great-grandfather kind leonidas, the pious baylor the blessed, and maegor the cruel"). Note that the antecedents in these examples are very long and certainly pose challenges to a mention detection system.

Next, we take a closer look at the top-performing system, the Emory system, in an attempt to understand why it works better than the other systems. First, there are more than 1700 links in the AMI test corpus that involve "well" (i.e., "we'll") and "were" (i.e., "we're") and are correctly identified by Emory but not the other systems. In fact, the large discrepancy in resolution performance between Emory and UTD on AMI can primarily be attributed to UTD's failure to even extract "well" and "were" as mentions (probably because of the missing apostrophe). Second, Emory appears to be better than the other systems at exploiting context to determine when two mentions are coreferent. Consider a lexical pair that appears frequently in the data such as ("I", "I"). While many instances of it are coreferent, there are also many instances that are not coreferent. The coreferent and non-coreferent instances can only be distinguished by factors such as distance and the surrounding context. While Emory and UTD achieve similar recall numbers, Emory achieves higher precision scores because of its ability to better exploit context to distinguish the coreferent and non-coreferent cases of a frequently-occurring lexical pair than UTD.

We also examine the erroneous links established by Emory. A major type of error involves links between nouns that are synonymous or semantically similar. Examples include ("contrast", "brightness") and ("the cash", "budget"). Another major type of error involves the frequently occurring lexical pairs discussed in the previous paragraph: while Emory is comparatively better than the other systems in exploiting context to distinguish the coreferent and non-coreferent instances, it is still far from being perfect in doing so. For instance, while it correctly identifies more than 1700 links involving "well" and "were", it incorrectly identifies more than 1500 links involving these two pronouns. Determining how to effectively exploit context to distinguish the coreferent instances from their non-coreferent counterparts is by no means trivial, but it is a problem that must be addressed in order to bring entity coreference resolvers to the next level of performance.

Table 8 shows examples of the most frequent gold coreferent pairs in the four test sets as well as the predictions made by the four systems on these pairs. Specifically, the first block shows the results of the five most frequent coreferent pairs. Perhaps not surprisingly, each pair involves two pronouns. The "count" column shows the number of times the two pronouns are coreferent in the test data. In addition, we show the number of times each system correctly predicts each pair as coreferent as well as the number of times each system incorrectly predicts each pair as coreferent. As can be seen, Emory and UTD establish a lot more correct links but also a lot more erroneous links than KU and DFKI. Moreover, the number of wrong links posited by Emory is considerably smaller than that by UTD. These results provide empirical support for our earlier claim that determining how to effectively exploit context to distinguish the coreferent instances of a frequent pair from their non-coreferent counterparts is a challenging but important problem for coreference researchers.

The second block shows results involving pairs in which one is a pronominal mention and the other is a non-pronominal mention, whereas the last block shows results involving pairs in which both mentions are non-pronominal. The first two blocks of results are similar in terms of the ob-

anaphor	antecedent	count	Emory		UTD		KU		DFKI	
			Correct	Wrong	Correct	Wrong	Correct	Wrong	Correct	Wrong
Most frequent pairs										
i	i	34534	31110	13677	32175	17729	8120	5420	29747	6515
i	you	8500	7053	3484	6866	7438	2697	2120	2607	7056
we	we	8148	5957	8829	8049	12885	1412	264	7858	12793
you	i	7472	6155	3778	6113	7676	2642	2131	3402	10122
you	you	4713	3996	1647	3962	4241	2304	1557	4148	16741
Most frequent (pronoun, non-pronoun) or (non-pronoun, pronoun) pairs										
they	people	114	88	55	107	141	26	8	23	36
it	the remote	109	23	24	59	181	9	0	1	0
i	william	96	0	0	0	0	0	0	0	0
people	them	93	78	13	90	15	6	0	0	0
people	they	89	61	75	85	73	17	11	0	0
Most frequent (non-pronoun, non-pronoun) pairs										
texas	here	19	0	0	0	0	0	1	0	0
texas	down here	18	0	0	0	0	0	0	0	0
the product	the remote	18	0	0	10	2	1	1	0	0
here	raleigh	15	0	0	15	41	0	0	0	0
childrens	children	12	6	3	6	3	2	3	0	0

Table 8: Entity coreference resolution: examples of frequent gold coreferent pairs and the results of the systems on them.

servations we can draw. The last block of results indicate how challenging it is to correctly establish links between two non-pronominal mentions.

3 Bridging Resolution

The shared task divides the evaluation of bridging resolution into two phases: (1) the *Predicted* phase, where a system needs to first identify all of the entity mentions that likely correspond to anaphors and antecedents, then perform bridging resolution on the predicted mentions; and (2) the *Gold* phase, which is essentially the same as the Predicted phase except that bridging resolution is performed on the given gold mentions.

In this section, we analyze the performance of the teams that participated in the bridging resolution track. The UTD team (Kobayashi et al., 2021) and the KU team (Kim et al., 2021) participated in both phases, whereas the INRIA team (Renner et al., 2021) only participated in the Gold phase. In other words, two teams participated in the Predicted phase, and three teams participated in the Gold phase. We will use their team name to refer to the bridging resolution systems they developed. To make it clear which phase a system was developed for, we will augment the team name with a superscript that encodes the phase. For instance, we will use UTD^P and UTD^G to refer to the systems the UTD team developed for the Predicted phase and the Gold phase respectively.

3.1 Anaphor Extraction

The "Recognition" rows in Table 9 show the official anaphor extraction results on each of the four test sets, where results are expressed in terms of recall (R), precision (P), and F-score. An anaphor is considered correctly detected if it has an exact match with a gold bridging anaphor in terms of boundary. Comparing the two systems developed for the Predicted phase, we see that KU^P beats UTD^P on three datasets, LIGHT, AMI, and Switchboard. Perhaps impressively, on these three datasets, KU^P outperforms UTD^P in terms of both precision and recall, showing its firm superiority over UTD^P. Note that KU^P is a pipelined system where anaphor extraction is performed as an explicit step prior to resolution, whereas UTD^P is a span-based system where the spans corresponding to anaphors are jointly learned as part of the resolution process. These results seem to suggest that better results can be achieved if one designs a model specifically for anaphor extraction.

Among the three Gold systems, UTD^G outperforms KU^G on all datasets, and KU^G in turn outperforms INRIA^G on all datasets. One difference between UTD^G and UTD^P is that the former has an explicit anaphor extraction component whereas the latter does not. The better anaphor extraction results achieved by UTD^G in comparison to KU^G could be therefore be attributed to the introduction of this anaphor extraction component, providing further empirical support for our earlier hypoth-

	LIGHT			AMI			Persuasion			Switchboard		
	P	R	F	P	R	F	P	R	F	P	R	F
					UTD^P							
Recognition	21.8	44.3	29.2	26.8	32.9	29.5	29.6	38.9	33.6	28.9	32.2	30.4
Resolution	10.4	21.2	14.0	12.1	14.8	13.3	19.3	25.3	21.9	14.5	16.1	15.3
					KU^P							
Recognition	28.8	54.7	37.7	32.3	38.4	35.1	20.9	60.4	31.1	29.7	43.4	35.3
Resolution	10.3	19.5	13.5	9.4	11.2	10.3	8.3	24.0	12.3	9.3	13.5	11.0
					UTD^G							
Recognition	34.7	40.7	37.5	37.0	42.2	39.4	43.0	52.1	47.1	37.7	50.9	43.3
Resolution	18.3	21.4	19.7	18.4	21.0	19.6	28.7	34.7	31.4	18.4	24.8	21.1
					KU^G							
Recognition	38.2	34.7	36.4	41.9	30.8	35.5	31.3	53.1	39.4	41.2	30.9	35.3
Resolution	17.5	15.9	16.7	18.1	13.3	15.3	14.9	25.3	18.8	21.4	16.1	18.3
					INRIA^G							
Recognition	34.8	11.8	17.6	34.1	14.4	20.2	46.7	17.0	24.9	34.2	24.9	28.8
Resolution	18.4	6.3	9.4	10.1	4.3	6.0	30.5	11.1	16.3	9.2	6.7	7.8

Table 9: Bridging resolution: official anaphor recognition and resolution results.

esis that better anaphor extraction results could be achieved via an anaphor extraction component. While one would generally expect to see better anaphor extraction performance in the Gold phase than in the Predicted phase, it is interesting to see that this is not necessarily the case for KU. Specifically, except on Persuasion where KU^G achieves considerably better performance than KU^P, the two systems achieve similar F-scores on the remaining three datasets. While their F-scores are similar, the recall and precision scores are not: KU^P has dramatically higher recall and substantially lower precision than KU^G. INRIA^P is roughly on par with the other systems in terms of precision, but its recall is much lower than the other systems: the best recall it achieves on any of the datasets is only 24.9%. This level of extraction performance will likely limit its resolution performance severely.

Tables 10 and 11 show each system's coverage of correct and wrong anaphors in the Predicted phase and the Gold phase respectively. The rows and columns in these tables can be interpreted in the same way as those in Table 2. As a quick reminder, "none" shows the percentage of gold anaphors that are not extracted by any of the systems, and the "x-only" columns show the percentage of correct/wrong anaphors extracted by system x and not any of the other systems.

Consider first the left half of Table 10, which shows each Predicted system's coverage of gold anaphors. As we can see, 40.4% of the gold anaphors are not extracted by any of the two systems. More specifically, 38.4% of the non-pronominal anaphors (the largest group), 50.7% of the 1st+2nd person pronouns (one of the second largest groups) and 67.6% of the 3rd person ungendered pronouns (the other second largest group) are not extracted by any of them. These results suggest that anaphor extraction in the Predicted phase is rather challenging. In terms of the coverage of gold anaphors, the "x-only" columns show that the two Predicted systems are quite different: while many of the anaphors extracted by KU^P are not extracted by UTD^P, there are also a number of anaphors that are extracted by UTD^P but not by KU^P.

The right half of Table 10 shows each Predicted system's coverage of mentions that are erroneously extracted as anaphors. As can be seen, KU^P extracts nearly half of the erroneously extracted anaphors, whereas the corresponding percentage for UTD^P is slightly lower (38.9%). The number of mistakes uniquely made by KU^P is larger than that by UTD^P on all but Groups 4 (possessive pronouns) and 7 (demonstrative pronouns). These results suggest that the two systems are very different from each other in terms of anaphor extraction.

Next, consider the three Gold systems in Table 11. Somewhat surprisingly, when the gold mentions are provided, the percentage of anaphors that cannot be extracted by any of the three systems increases for all but two groups (Group 10 (non-pronominal mentions) and Groups 6 (indefinite pronouns)) in comparison to the corresponding percentages in the Predicted phase. The remaining

Group	Coverage of gold anaphors %	count	none	T-only	K-only	Coverage of wrong anaphors %	count	T-only	K-only
Overall	100.0	2526	40.4	14.5	25.1	100.0	4863	38.9	49.5
1	2.8	71	50.7	8.5	26.8	9.4	459	14.6	80.4
2	0.0	1	100.0	0.0	0.0	0.2	10	0.0	100.0
3	2.8	71	67.6	2.8	25.4	3.3	159	8.2	85.5
4	0.9	23	39.1	26.1	30.4	1.1	55	58.2	30.9
5	0.0	0	100.0	0.0	0.0	0.1	3	33.3	66.7
6	1.8	46	26.1	17.4	30.4	2.0	97	58.8	34.0
7	1.9	49	73.5	0.0	22.4	1.9	94	6.4	91.5
8	0.0	1	100.0	0.0	0.0	0.1	4	25.0	75.0
9	1.3	32	62.5	9.4	21.9	0.8	39	33.3	64.1
10	88.4	2232	38.4	15.3	25.0	81.1	3943	43.1	43.8

Table 10: Bridging resolution (Predicted phase): coverage of correct and wrong anaphors.

Group	Coverage of gold anaphors %	count	none	T-only	K-only	I-only	Coverage of wrong anaphors %	count	T-only	K-only	I-only
Overall	100.0	2526	35.6	20.2	12.3	3.5	100.0	3338	37.5	26.4	12.3
1	2.8	71	56.3	11.3	5.6	8.5	8.9	297	26.3	37.0	19.2
2	0.0	1	100.0	0.0	0.0	0.0	0.1	5	0.0	60.0	40.0
3	2.8	71	73.2	5.6	7.0	2.8	3.5	117	47.0	31.6	12.8
4	0.9	23	56.5	13.0	0.0	0.0	1.0	35	48.6	5.7	31.4
5	0.0	0	100.0	0.0	0.0	0.0	0.0	1	0.0	100.0	0.0
6	1.8	46	17.4	19.6	15.2	2.2	1.9	62	30.6	40.3	9.7
7	1.9	49	83.7	2.0	12.2	0.0	1.0	35	20.0	71.4	8.6
8	0.0	1	100.0	0.0	0.0	0.0	0.1	3	0.0	0.0	100.0
9	1.3	32	75.0	6.2	3.1	3.1	0.6	20	40.0	55.0	5.0
10	88.4	2232	32.2	21.7	12.9	3.5	82.8	2763	38.7	24.1	11.3

Table 11: Bridging resolution (Gold phase): coverage of correct and wrong anaphors.

columns provide suggestive evidence that the three systems are quite different from each other in terms of anaphor extraction. Specifically, among the gold anaphors, 20.2% are extracted only by UTD^G, 12.3% only by KU^G, and 3.5% only by INRIA^G, and among the erroneously extracted anaphors, 37.5% are only extracted by UTD^G, 26.4% only by UTD^G, and 12.3% only by INRIA^G.

3.2 Resolution

The "Resolution" rows in Table 9 show the official resolution results on each of the four test sets, where results are expressed in terms of recall (R), precision (P), and F-score at the *entity* level. In other words, an anaphor is considered correctly resolved if it is resolved to its antecedent or any preceding mention that is coreferent with its antecedent. Comparing the two Predicted systems, we see that UTD^P beats KU^P on all datasets in terms of recall, precision, and F-score. Given that UTD^P's anaphor extraction performance is poorer than that of KU^P, its superior resolution results can be attributed solely to better resolution and not better anaphor extraction. We speculate that KU^P's poorer resolution performance can be attributed to its attempt to establish links, many of which are wrong, more aggressively than UTD^P.

Among the three Gold systems, UTD^G outperforms KU^G on all datasets in terms of F-score, and with the exception of its precision on Switchboard, it outperforms KU^G on both recall and precision on all datasets. It is worth noting that both Gold systems outperform their Predicted counterparts on all datasets, which is consistent with our expectation that the Gold setting is easier than the Predicted setting. INRIA^G's performance is much lower than the other two systems in terms of both recall and precision, but primarily because of recall. We believe that its low recall can be attributed in part to its fairly poor anaphor extraction performance.

To gain a better understanding of how these systems perform w.r.t. identifying difficult vs. easy links, we divide the bridging pairs based on how hard we believe it is to resolve them. Specifically, we partition the bridging pairs into five groups: same string (the two mentions are the same string), same head (the two mentions have the same head), same head lemma (the two mentions have the same lemma head), word overlap (the two mentions have at least one content word overlap), and other (pairs

(a) Coverage of correct links

Group	all		none		T		K		T-only		K-only	
	%	cnt	%	cnt	%	cnt	%	cnt	%	cnt	%	cnt
Overall	100.0	2526	77.4	1954	14.4	364	12.1	306	10.5	266	8.2	208
Same string	1.1	28	89.3	25	10.7	3	0.0	0	10.7	3	0.0	0
Same head	18.8	475	58.5	278	32.0	152	17.7	84	23.8	113	9.5	45
Same head lemma	6.6	167	61.1	102	30.5	51	15.6	26	23.4	39	8.4	14
Word overlap	4.2	105	79.0	83	13.3	14	13.3	14	7.6	8	7.6	8
Other	69.3	1751	83.7	1466	8.2	144	10.4	182	5.9	103	8.1	141

(b) Coverage of wrong links

Group	all		T		K		T-only		K-only	
	%	cnt	%	cnt	%	cnt	%	cnt	%	cnt
Overall	100.0	6604	44.8	2961	57.6	3807	42.4	2797	55.2	3643
Same string	7.0	462	54.1	250	45.9	212	54.1	250	45.9	212
Same head	16.0	1055	74.3	784	31.2	329	68.8	726	25.7	271
Same head lemma	4.3	287	71.1	204	39.0	112	61.0	175	28.9	83
Word overlap	5.2	346	46.5	161	55.8	193	44.2	153	53.5	185
Other	67.4	4454	35.1	1562	66.5	2961	33.5	1493	64.9	2892

Table 12: Bridging resolution (Predicted phase): coverage of correct and wrong links.

that have no lexical overlap).

Two points deserve mention. First, the groups are listed in ascending order of resolution difficulty: intuitively, a pair of mentions having the same head lemma is easier to resolve than one that does not have any lexical overlap, for instance. Second, if a pair belongs to a group (e.g., same head lemma), then it should also belong to all subsequent groups (i.e., word overlap and other), but since we are *partitioning* the pairs, we will assign a pair to only the first group that is applicable to it.

Tables 12a and 13a show the results for the Predicted systems and the Gold systems respectively. The rows correspond to the five groups. The columns can be interpreted in the same way as those in Table 7a. "all" expresses the size of a group in terms of the number of pairs it covers and the corresponding percentage. "none" shows the number and percentage of pairs that are not resolved by any of the systems. The "T", "K", and "I" columns show the number and percentage of pairs correctly resolved by the individual systems, whereas the "x-only" columns show the number and percentage of pairs that can be resolved by system x and not any of the other systems.

We begin by noting that the percentage of anaphors that belong to the "same string" group is much smaller in bridging than in coreference. This is understandable: this group contains pairs in which the mentions are lexically identical. While many coreferent mentions are lexically identical, relatively few bridging pairs are composed of lexically identical mentions. In addition, approximately only 30% of the links (i.e., those that connect the pairs in the first four groups) can be established using string-matching facilities. This makes bridging resolution more challenging than entity coreference resolution.

Next, consider the Predicted systems. As we can see in the "overall" row, the overall resolution performance of KU^P is worse than that of UTD^P. However, this by no means implies that UTD^P is better than KU^P in all categories. Generally, UTD^P performs much better than KU^P on the easier categories, and as we move down the table, the performance gap between the two systems continues to shrink, to the point that KU^P starts to outperform UTD^P on "other", the most difficult category. Overall, these results suggest that while UTD^P is better than KU^P at resolving the easier-to-resolve pairs, the reverse is true when it comes to resolving the difficult-to-resolve pairs. Considering the results in the "x-only" columns, we see that the links recalled by the two systems are largely different from each other. Finally, considering the results in the "none" column, we note that 77.4% of the links are not recalled by any of the two systems. Even for simpler categories such as "same head" and "same head lemma", around 60% of the pairs are not recalled. These results provide suggestive evidence that the Predicted setting is indeed very challenging.

Consider the Gold systems. A few points deserve mention. First, the performance differences that

(a) Coverage of correct links

Group	all		none		T		K		I		T-only		K-only		I-only	
	%	cnt	%	cnt	%	cnt	%	cnt	%	cnt	%	cnt	%	cnt	%	cnt
Overall	100.0	2526	70.2	1773	19.3	488	13.5	342	5.3	134	12.5	315	7.7	195	1.9	49
Same string	1.1	28	89.3	25	0.0	0	0.0	0	10.7	3	0.0	0	0.0	0	10.7	3
Same head	18.8	475	47.4	225	42.5	202	19.8	94	7.6	36	27.2	129	7.8	37	1.5	7
Same head lemma	6.6	167	46.1	77	37.7	63	27.5	46	10.8	18	21.6	36	12.6	21	1.2	2
Word Overlap	4.2	105	64.8	68	21.9	23	17.1	18	5.7	6	16.2	17	8.6	9	1.9	2
Other	69.3	1751	78.7	1378	11.4	200	10.5	184	4.1	71	7.6	133	7.3	128	2.0	35

(b) Coverage of wrong links

Group	all		T		K		I		T-only		K-only		I-only	
	%	cnt	%	cnt	%	cnt	%	cnt	%	cnt	%	cnt	%	cnt
Overall	100.0	5400	49.5	2675	34.9	1883	23.0	1244	43.2	2333	30.2	1629	19.6	1060
Same string	7.7	414	77.1	319	11.8	49	17.6	73	70.5	292	11.8	49	11.1	46
Same head	15.7	847	74.0	627	29.0	246	11.1	94	61.4	520	19.0	161	6.6	56
Same head lemma	4.6	250	71.6	179	32.0	80	11.2	28	58.0	145	21.6	54	6.4	16
Word overlap	7.7	416	34.9	145	31.5	131	39.7	165	30.3	126	27.4	114	37.5	156
Other	64.3	3473	40.5	1405	39.6	1377	25.5	884	36.0	1250	36.0	1251	22.6	786

Table 13: Bridging resolution (Gold phase): coverage of correct and wrong links.

we have observed above between UTD^P and KU^P are also applicable to UTD^G and KU^G, except that UTD^G outperforms KU^G on all categories. In other words, UTD^G manages to do better than KU^G on the difficult categories. Second, INRIA^G underperforms UTD^G and KU^G on all but the easiest group, "same string". Third, considering the results in the "x-only" columns, we see that the links recalled by UTD^G and KU^G are quite different, but the links recalled by INRIA^G are for the most part similar to those recalled by the other two systems. Finally, considering the results in the "none" column, we see that 70.2% of the links are not recalled by any of the three systems, which is smaller than the corresponding percentage in the Predicted setting. In fact, the percentage associated with nearly every group in the Gold phase is smaller than the corresponding percentage in the Predicted phase. This again provides suggestive evidence that the Gold setting is indeed less challenging than the Predicted setting.

While Tables 12a and 13a focus on the extraction of gold pairs, Tables 12b and 13b focus on the extraction of wrong pairs (i.e., wrong links). Specifically, in Table 12b, we take the union of the set of wrong pairs extracted by the two Predicted systems and compute statistics based on the resulting set, which we refer to as S. The columns in Table 12b can be interpreted in the same way as those in Table 12a. Table 13b is essentially the same as Table 12b except that it shows the results of the three Gold systems.

Consider first the Predicted systems (Table 12b). As can be seen, 6604 erroneous links are established by the two systems, of which 44.8% are established by UTD^P and 57.6% by KU^P; in addition, 42.4% of these erroneous links are only established by UTD^G and 55.2% are only established by UTD^P. Comparing Tables 12a and 12b, we see that each system identifies a lot more erroneous links than correct links. Moreover, while Table 12a shows that UTD^P performs better than KU^P on the easy categories and worse than it on the harder categories, Table 12b shows the reverse trend. Specifically, UTD^P extracts more erroneous pairs that belong to the easy categories than KU^P, whereas KU^P extracts more erroneous pairs that belong to the harder categories than UTD^P. Finally, considering the "x-only" columns, we see that there is very little overlap in the erroneous links predicted by the two systems.

Next, consider the Gold systems (Table 13b). We can see that 5400 erroneous links are established by the three systems, which is smaller than the corresponding number in Table 12b. This again suggests that the provision of gold mentions has made the task somewhat easier. Of these 5400 erroneous links, 49.5% are established by UTD^G, 34.9% by KU^G, and 23.0% by INRIA^G; in addition, 43.2% of these erroneous links are only established by UTD^G, 30.2% are only established by UTD^G, and 19.6% are only established by INRIA^G. Some of the observations we made on the Predicted results in Table 12b are applicable to the Gold results. For

relation/category	Recalled anaphor	antecedent	T^P	K^P	T^G	K^G	I^G	Missed anaphor	antecedent
Part-whole	these waters	the sea	✓	✓	✓	✓	✓	a bird	a beak
Is-a	a parent	the father	✓		✓			a chrysler	car
Instance-of (same head)	a car	cars	✓		✓			a college	colleges
Instance-of (diff heads)	a highway	the road		✓		✓		humans	i
Related/Associated	any amount	your payment	✓	✓	✓	✓	✓	sound	volume
Number	fifteen	seventeen	✓	✓		✓		fifty	the tenth
Pronoun pairs	we	you		✓				we	we
Pro, non-pro pairs	they	texas				✓		people	they
Demonstrative pronouns	this	an adventure		✓		✓		conceptual design	this

Table 14: Bridging resolution: examples of gold bridging pairs that belong to different relations/categories.

instance, UTD^G extracts a lot more erroneous pairs that belong to the easy categories than KU^P. What is different is that UTD^G extracts more erroneous pairs than KU^G on the hard categories as well, even though the difference in the number of erroneous pairs they extract is smaller as the hardness level increases. $INRIA^G$ generally extracts fewer erroneous pairs than the other two systems, but it extracts more erroneous pairs in the "word overlap" group than the other systems. Finally, considering the "x-only" columns, we see that the three systems are quite different from each other in terms of their prediction of erroneous pairs.

3.3 Discussion

In this subsection, we manually analyze the pairwise links that are correctly and incorrectly established by the participating systems.

A bridging relation in this data typically involves two mentions where one is a specific instance of a generic concept referred to by the other via the use of semantic relations such as *set-subset*, *part-whole*, and *is-a*. There are no noticeable differences between the Predicted systems and the Gold systems in terms of the *kind* of semantic relations they extract. Specifically, both the Predicted systems and the Gold systems are able to extract a variety of semantic relations as bridging relations, such as *is-a* (e.g., ("an eagle", "bird"), ("a light blue", "standard color")), set-subset (e.g., ("I", "we"), ("charities", "Save the Children")) and *part-whole* (e.g., ("the engine", "the car"), ("New Orleans", "the United States")). In addition, both are able to extract less well-defined relations in which one mention is simply associated with or is a specific instance of the other (e.g., ("one chip", "two"), ("a child", "kid"), ("people", "customers"), ("it", "a new one, the phone"), ("the place", "a restaurant"), ("some", "animals"), ("That", "a special flower to show you"), ("dresses", "Levi's), ("people who have killed police officers", "murderers")). Additional examples that are successfully recalled by the systems are shown in Table 14. Specifically, the "Recalled" column shows successfully recalled pairs that are instances of different semantic relations or categories as well as the system(s) that identified these pairs. As discussed before, while some relations can be identified via string-matching facilities (e.g., ("the function", "a desired function"), ("two batteries", "battery")), the majority of them cannot.

In addition, we do not observe any noticeable differences between the systems submitted by different teams in terms of the kinds of semantic relations they extract. As discussed before, the UTD systems achieve better resolution results than the KU systems because (1) the UTD systems are more conservative in positing bridging links between two mentions than the KU systems, and (2) the UTD systems focus more on the relations that can be identified via string-matching facilities, which presumably are easier to identify.

Examining the links that are missed by all of the systems, we do not find any noticeable differences between the kinds of semantic relations that exist in the correctly extracted pairs and those that exist in the pairs that fail to be extracted. Table 14 shows several gold pairs that are not extracted by any of the systems in the "Missed" column. While a deeper analysis is needed in order to understand why some instances of a particular semantic relation are being extracted and others are not, we speculate that the distance between the two mentions involved, their surrounding contexts, and whether a given mention pair is seen in the training data as having a bridging relation may have played a role.

While it is encouraging to see from Table 14 that the systems can successfully recall pairs that belong to the "Other" (i.e., most difficult-to-resolve)

category, a large percentage of bridging links in this category are still not extracted by any of the systems. To improve system recall for these difficult cases, existing work has explored the use of information extracted from manually constructed resources such as knowledge graphs (Pandit et al., 2020) as well as bridging pairs (Hou, 2018) and bridging-annotated data (Hou, 2020) automatically constructed using lexico-syntactic patterns. These and other ideas (see Kobayashi and Ng (2020) for an overview) could be explored to improve the recall of the participating systems. Note, however, that these manually and automatically constructed resources are not likely to be helpful for resolving bridging links that involve pronouns as well as nouns that are semantically poor (e.g., "here"). Currently, demonstrative pronouns such as "this" and "that" have a resolution recall of 14.2% and 4.5% respectively, and the word "here" has a resolution recall of 6.3%. Given that these anaphors cannot be resolved via string matching, the only way to resolve them is to exploit the contexts in which they appear. While the mention representations acquired by existing mechanisms are supposed to be contextualized, the contextual information encoded in them is arguably quite limited and insufficient as far as making accurate linking decisions is concerned. Hence, learning effective context representations is a key challenge for state-of-the-art bridging resolvers. As discussed earlier, learning effective context representations is also an issue surrounding state-of-the-art entity coreference resolvers. However, we believe that this issue is likely to be more challenging for bridging resolution than for entity coreference. The reason is that determining whether two mentions are *associated* based on context is in general a lot more challenging than determining whether they refer to the same entity.

4 Discourse Deixis Resolution

The shared task divides the evaluation of discourse deixis resolution into two phases: (1) the *Predicted* phase, where a system needs to first identify all of the entity mentions that likely correspond to anaphors and antecedents, then perform discourse deixis resolution on the predicted mentions; and (2) the *Gold* phase, which is essentially the same as the Predicted phase except that the mentions corresponding to anaphors are to be extracted from the given gold mentions.

In this section, we analyze the performance of the teams that participated in the discourse deixis resolution track. The UTD team (Kobayashi et al., 2021) participated in both phases, whereas the DFKI team (Anikina et al., 2021) participated in the Predicted phase only. In other words, two teams participated in the Predicted phase, and one team participated in the Gold phase. As in bridging, we will use their team name to refer to the discourse deixis resolution systems they developed, augmenting the team name with a superscript that encodes the phase the system was developed for.

4.1 Mention Extraction

Mention extraction results of the four test sets, which are expressed in terms of R, P, and F, are shown in the "Overall" row of Table 15. As in the other tracks, a mention is considered correctly extracted if it has an exact match with a gold mention in terms of boundary.

Consider first the two systems developed for the Predicted phase, UTD^P and DFKI^P. DFKI^P achieves a much higher recall (9–17% points) than UTD^P on three of the four datasets, and on the remaining dataset (LIGHT), the two systems achieve comparable recall. These results suggest that DFKI^P is a lot more aggressive in extracting mentions than UTD^P. The high recall scores achieved by DFKI^P, however, come at the expense of precision. As can be seen, DFKI^P's precision scores are substantially lower than UTD^P's, with a difference of 37–42% points. Consequently, the mention extraction F-scores achieved by DFKI^P are also lower than those achieved by UTD^P: there is a 12–22% point difference in F-score between the two systems.

Next, consider the two UTD systems, UTD^P and UTD^G. The performance difference between these two systems is less than that between the two systems for the Predicted phase. In terms of F-score, while UTD^G outperforms UTD^P by nearly 12% points on Persuasion, the two differ from each other by only 2–3% points on the remaining datasets. The difference between their recall and precision values, however, provides some evidence that they may not be as similar to each other as their F-scores suggest. Specifically, UTD^G's mention extraction system seems to be recall-oriented: on three of the four datasets, UTD^G has a much higher recall (6–26% points) but a much lower precision (6–16% points) than UTD^P.

To better understand whether these systems dif-

		LIGHT			AMI			Persuasion			Switchboard		
		P	R	F	P	R	F	P	R	F	P	R	F
Overall	UTD^G	66.2	49.0	56.3	54.5	45.2	49.4	61.8	67.3	64.4	48.2	61.8	54.2
	UTD^P	65.2	46.9	54.5	60.2	39.1	47.4	72.3	41.6	52.8	64.4	42.2	51.0
	DFKI^P	25.2	44.3	32.1	18.5	56.3	27.8	31.5	58.4	40.9	27.7	51.3	36.0
Anaphor	UTD^G	–	65.0	–	–	61.9	–	–	77.2	–	–	74.9	–
	UTD^P	–	73.8	–	–	64.4	–	–	65.9	–	–	71.1	–
	DFKI^P	–	56.2	–	–	81.4	–	–	69.1	–	–	68.1	–
Antecedent	UTD^G	–	37.5	–	–	32.9	–	–	58.9	–	–	52.4	–
	UTD^P	–	27.7	–	–	20.5	–	–	21.2	–	–	21.5	–
	DFKI^P	–	35.7	–	–	37.9	–	–	49.3	–	–	39.4	–

Table 15: Discourse deixis resolution: mention extraction results.

fer in terms of how well they extract anaphors and antecedents, we also show in the last two rows of Table 15 their results on anaphor and antecedent extraction. Since each mention in the test sets is annotated as "anaphor" or"antecedent", we can easily compute recall. However, since the systems did not label each of the extracted mentions as "anaphor" or "antecedent" in the outputs, we cannot compute precision. As can be seen, DFKI^P extracts more mentions as antecedents and anaphors than UTD^P on all datasets, with the exception on LIGHT and Switchboard, where UTD^P achieves better recall on anaphor extraction. Comparing UTD^P and UTD^G, we can see that while UTD^G lags behind UTD^P by 3–9% points in anaphor extraction on LIGHT and AMI, UTD^G achieve superior mention extraction performance to UTD^P in the remaining cases. In particular, the difference between their recall in antecedent extraction is much bigger than that in anaphor extraction.

To gain additional insights into the difference between the systems w.r.t. mention extraction, we show in Table 16a their performance on extracting the five gold mentions that occur most frequently on the four test sets combined. More specifically, in the "mention" column, "overall" shows the results aggregated over all gold mentions, "The rest" aggregates the results of all but the top five gold mentions, and the remaining rows show the results of each of the top five gold mentions. The "%" and "count" columns show the percentage and number of gold mentions that belong to each category. The remaining columns can be interpreted in the same way as those in Table 2. For instance, "none" shows the percentage of gold mentions that cannot be extracted by any of the three systems, UTD^G shows the percentage of gold mentions extracted by UTD^G, and UTD^G-only shows the percentage of gold mentions that are extracted by UTD^G but not by the other two systems.

Perhaps not surprisingly, the most frequent categories of mentions are all anaphor categories. Specifically, the most frequent category is "that", accounting for 29.3% of the gold mentions. This is followed by "it" (6.9%) and "this" (1.8%). These three categories of anaphors account for nearly 38% of the gold mentions in the test data. From the "none" column, we see that the worst-performing category is "The rest". This should perhaps not be surprising either: the majority of the mentions in this category are antecedents, which may not have appeared in the training data at all. In addition, 55.8% of the anaphor "it" were missed by the systems. We believe that this can be attributed to two reasons. First, DFKI^P gave up on handling "it". Second, we speculate that it is challenging to determine whether "it" is used deictically: while "this", "that" and "it" can all be used as coreferent anaphors, bridging anaphors, and deictic expressions, it is more likely for "it" to be a coreferent or bridging anaphor than a discourse deixis compared to "this" and "that". Moving on to the performance of the systems (UTD^G, UTD^P, and DFKI^P), we see that the systems are indeed better at extracting "this" and "that" than "it". Note that DFKI^P's recall scores on the anaphors that the system is able to handle are substantially higher than those of the UTD systems. In particular, considering the "x-only" columns, we see that there are many instances of these anaphors that are extracted only by DFKI^P.

While Table 16a focuses on gold mention extraction, Table 16b focuses on the extraction of erroneous mentions. Specifically, we take the union of the set of erroneous mentions extracted by the three resolvers and compute statistics based on the

(a) Coverage of gold mentions

mention	%	count	none	UTD^G	UTD^P	DFKI^P	UTD^G-only	UTD^P-only	DFKI^P-only
Overall	100.0	1371	65.0	30.4	28.9	29.5	1.7	0.8	2.4
that	29.3	402	2.2	86.6	85.3	92.3	1.2	0.7	6.0
it	6.9	95	55.8	36.8	32.6	0.0	11.6	7.4	0.0
this	1.8	25	0.0	76.0	60.0	100.0	0.0	0.0	20.0
which	0.7	10	10.0	50.0	30.0	70.0	10.0	0.0	40.0
the same	0.3	4	25.0	75.0	25.0	0.0	50.0	0.0	0.0
The rest	60.9	835	99.0	0.8	0.4	0.1	0.5	0.1	0.0

(b) Coverage of wrong mentions

mention	%	count	UTD^G	UTD^P	DFKI^P	UTD^G-only	UTD^P-only	DFKI^P-only
Overall	100.0	1202	43.3	17.4	61.6	29.8	4.1	49.9
that	43.3	520	25.8	23.5	90.6	2.7	2.1	67.7
this	6.3	76	17.1	23.7	89.5	3.9	2.6	71.1
it	4.5	54	61.1	81.5	0.0	18.5	38.9	0.0
which	3.5	42	9.5	11.9	95.2	2.4	2.4	81.0
that way	0.4	5	80.0	60.0	0.0	40.0	20.0	0.0
The rest	42.0	505	65.7	3.4	31.9	65.0	2.6	31.7

Table 16: Discourse deixis resolution: per-anaphor mention extraction results.

resulting set, which we refer to as S. The columns in Table 16b can be interpreted in the same way as those in Table 16a. For instance, UTD^G, UTD^P, and DFKI^P show the percentage of mentions in S extracted by each of the systems, and the "x-only" columns show the percentage of mentions in S extracted by exactly one of the systems.

A few points deserve mention. First, the five most frequently occurring categories of mentions that are erroneously extracted are likely mentions that are incorrectly posited by the systems as discourse deixis, as the top four categories are the same as the top four categories of gold mentions. The erroneously extracted antecedents will likely appear in the "The rest" category. Second, DFKI^P covers more than 90% of the mistakes made for the top categories of anaphors that it can handle (i.e., "that", "this", "which"), suggesting that the system is very aggressive in positing the occurrences of these words as anaphors. In contrast, it only extracts one-third of the mentions in the "The rest" category, which, as noted above, should mostly contain erroneously extracted antecedents, while UTD^G and UTD^P extract two-thirds and 3% of the erroneously extracted antecedents respectively. These results imply that UTD^P is a lot more cautious in positing mentions as antecedents compared to the other two systems, while UTD^G is the most liberal in positing mentions as antecedents. These differences can also be observed when considering the "x-only" columns.

4.2 Resolution

A discourse deixis resolver is expected to output clusters, each of which contains a deixis and all of its antecedents. As far as scoring is concerned, discourse deixis resolution is viewed as a generalized case of event coreference, and hence the scorer used for scoring entity coreference chains is used to score the output of a discourse deixis resolver.

Table 17 shows the official results obtained using the official scorer. Again, the results are expressed in terms of MUC, B^3, and CEAF$_e$ R, P, and F, as well as the CoNLL score.

A few points deserve mention. First, there is a clear difference in performance between the two systems developed for the Predicted phase, UTD^P and DFKI^P, in terms of CoNLL F-score, where the scores achieved by UTD^P almost double those achieved by DFKI^P. In comparison, the difference in the CoNLL score between UTD^G and UTD^P is smaller: less than 2% points on LIGHT and AMI, 12.5% points on Persuasion, and 5.0% points on Switchboard. We speculate that the particularly large difference in performance between the two systems on Persuasion can be attributed in part to mention extraction, where UTD^G has a considerably higher mention extraction F-score on this dataset than UTD^P (64.4 vs. 52.8). Second, DFKI^P's MUC recall scores are lower than the other systems for all but the Switchboard dataset. This implies that DFKI^P is not able to recall as many links as the other systems. However, a closer

	MUC			B^3			$CEAF_e$			
	P	R	F	P	R	F	P	R	F	CoNLL
					LIGHT					
UTD^G	49.0	30.0	37.2	56.3	39.1	46.2	51.7	42.9	46.9	43.4
UTD^P	44.6	31.2	36.8	56.2	37.0	44.6	55.3	40.5	46.7	42.7
$DFKI^P$	9.2	12.5	10.6	21.1	31.3	25.2	19.3	42.1	26.5	20.8
					AMI					
UTD^G	44.6	21.2	28.7	49.7	34.6	40.8	39.6	43.0	41.2	36.9
UTD^P	45.5	21.2	28.9	52.4	29.5	37.8	44.9	35.1	39.4	35.4
$DFKI^P$	7.5	16.9	10.4	14.5	39.1	21.2	12.6	49.5	20.1	17.2
					Persuasion					
UTD^G	53.3	45.5	49.1	54.9	55.7	55.3	46.0	59.3	51.8	52.1
UTD^P	45.5	20.3	28.1	64.9	30.2	41.2	61.0	41.8	49.6	39.6
$DFKI^P$	14.7	14.6	14.7	26.9	37.2	31.2	16.9	46.7	24.8	23.6
					Switchboard					
UTD^G	39.4	31.2	34.8	41.6	48.5	44.8	33.7	55.0	41.8	40.4
UTD^P	35.2	21.3	26.5	52.3	30.4	38.5	50.5	34.9	41.3	35.4
$DFKI^P$	14.2	21.3	17.1	22.6	37.2	28.1	18.9	42.8	26.2	23.8

Table 17: Discourse deixis resolution: official resolution results.

	LIGHT			AMI			Persuasion			Switchboard		
	F	ns. F	s. F	F	ns. F	s. F	F	ns. F	s. F	F	ns. F	s. F
UTD^G	43.4	43.9	7.7	36.9	35.2	8.4	52.1	52.0	4.2	40.4	36.6	9.5
UTD^P	42.7	47.0	2.2	35.4	37.9	2.6	39.6	42.1	1.0	35.4	39.1	2.0
$DFKI^P$	20.8	17.1	6.9	17.2	15.3	3.4	23.6	22.8	2.1	23.8	22.6	3.8

Table 18: Discourse deixis resolution: results on singleton and non-singleton cluster identification.

examination of the results reveals that precision seems to play a bigger role in the observed performance difference between $DFKI^P$ and the other systems: $DFKI^P$'s precision scores are generally very poor. We speculate that this can be attributed to the fact that the system is overly aggressive in positing "that", "this", and "which" as anaphors and attempts to resolve them, which in turn yields a lot of erroneous links. Third, comparing UTD^G and UTD^G, we see that the considerably better results achieved by UTD^G on Persuasion can be attributed to not only mention extraction but also resolution, as reflected in the 25.2% point difference in MUC recall between the two systems.

To further our understanding of how the systems perform w.r.t. non-singleton cluster and singleton cluster identification, we show the corresponding results in Table 18, whose rows and columns can be interpreted in the same manner as those in Table 5. Comparing the two Predicted systems, we see that while $DFKI^P$ is considerably worse than UTD^P in non-singleton cluster identification, it performs slightly and consistently better than UTD^P in singleton cluster identification. This again could be attributed to its being aggressive in extracting anaphors. Comparing the two UTD systems, we see that UTD^G is always better than UTD^P in singleton cluster identification, but UTD^P is better than UTD^G in non-singleton cluster identification on all but the Persuasion dataset. These results further reveal why UTD^G performs substantially better than UTD^P on Persuasion: UTD^G are better than UTD^P in both non-singleton and singleton cluster identification on Persuasion.

Next, we measure system performance, specifically link identification performance, at the pairwise level. For each non-singleton cluster in the gold output, we extract every pair of mentions in the cluster. We similarly extract all the pairs from the non-singleton clusters produced by each of the three systems. The recall, precision, and F-scores in Table 19 are computed based on the pairwise links in these sets.

From the "Overall" row in Table 19, we can see that 504 pairs can be extracted from the gold clusters of the four test sets. As we can see, all systems have higher recall than precision. Comparing the two Predicted systems, we see that while

anaphor	%	count	UTD^G			UTD^P			DFKI^P		
			P	R	F	P	R	F	P	R	F
Overall	100.0	504	25.0	37.7	30.1	17.2	26.2	20.7	9.4	23.6	13.5
that	68.1	343	30.1	47.8	36.9	21.4	33.5	26.1	11.8	33.5	17.5
it	17.7	89	12.7	16.9	14.5	9.1	13.5	10.9	0.0	0.0	0.0
this	4.4	22	18.9	31.8	23.7	10.5	18.2	13.3	2.9	13.6	4.8
which	2.0	10	20.0	30.0	24.0	0.0	0.0	0.0	1.9	10.0	3.1
the same	0.8	4	0.0	0.0	0.0	25.0	25.0	25.0	0.0	0.0	0.0
The rest	7.1	36	2.5	2.8	2.6	0.0	0.0	0.0	0.0	0.0	0.0

Table 19: Discourse deixis resolution: per-anaphor resolution results at the pairwise level.

(a) Coverage of correct links

anaphor	%	count	None	UTD^G	UTD^P	DFKI^P	UTD^G-only	UTD^P-only	DFKI^P-only
Overall	100.0	504	46.0	37.7	26.2	23.6	14.3	8.3	6.3
that	68.1	343	31.8	47.8	33.5	33.5	16.0	9.3	8.7
it	17.7	89	76.4	16.9	13.5	0.0	10.1	6.7	0.0
this	4.4	22	45.5	31.8	18.2	13.6	22.7	13.6	9.1
which	2.0	10	70.0	30.0	0.0	10.0	20.0	0.0	0.0
the same	0.8	4	75.0	0.0	25.0	0.0	0.0	25.0	0.0
The rest	7.1	36	97.2	2.8	0.0	0.0	2.8	0.0	0.0

(b) Coverage of wrong links

anaphor	%	count	UTD^G	UTD^P	DFKI^P	UTD^G-only	UTD^P-only	DFKI^P-only
Overall	100.0	1330	19.5	20.2	69.0	13.7	15.6	63.9
that	69.5	925	21.8	21.1	68.1	14.5	15.6	61.2
this	7.8	104	14.4	15.4	78.8	9.6	10.6	75.0
it	5.0	67	43.3	64.2	0.0	35.8	56.7	0.0
which	4.4	59	8.5	16.9	74.6	8.5	16.9	74.6
the	1.7	22	0.0	0.0	100.0	0.0	0.0	100.0
The rest	11.5	153	5.9	2.6	91.5	5.9	2.6	91.5

Table 20: Discourse deixis resolution: per-anaphor resolution results at the pairwise level.

DFKI^P achieves lower overall recall and precision than UTD^P, the difference in their recall scores is comparatively smaller than the difference in their precision scores. In particular, the two systems achieve the same recall in the resolution of "that", the most frequent anaphor, and the slightly lower overall recall achieved by DFKI^P can be largely attributed to its decision of not resolving "it" and some other anaphors. Comparing the two UTD systems, we see that UTD^G achieves better recall and precision than UTD^P in resolving the top anaphors. Perhaps more interesting, while the two Predicted systems cannot resolve any of the anaphors in "The rest" category, UTD^G manages to achieve a non-zero F-score on this category, though precision and recall are both low.

Table 20a shows each system's coverage of gold pairs. The rows can be interpreted in the same way as those in Table 19, whereas the columns can be interpreted in the same way as those in Table 7a. As a quick reminder, "none" shows the percentage of gold pairs that are not extracted by any of the three systems; "UTD^G", "UTD^P", and "DFKI^P" show the percentage of gold pairs extracted by each of the three systems; and the "x-only" columns show the percentage of gold pairs that are extracted only by system x.

A few points deserve mention. First, consider the "none" results. As can be seen, 46.0% of the links are not recovered by any of the three systems. In particular, 97.2% of the links involving the anaphors in the "The rest" category are not recovered. This is followed by "it", where 76.4% of the links involving "it" are not recovered. In contrast, the recovery rate is higher for "that", probably because of the larger representation of links involving "that" in the training set. Second, consider the "x-only" results, which suggest that the three systems are more different from each other than we may think. Examining the "that" links, we see that 16% are only identified by UTD^G, 9.3% only by UTD^P, and 8.7% only by DFKI^P. Similarly for "this": 22.7% are only identified by UTD^G, 13.6% only by UTD^P, and 9.1% only by DFKI^P.

	0	1	2	3	4	5	6	7	8	9	10	>10
Distribution of links over sentence distances												
Gold	90	216	98	49	21	9	4	2	0	1	3	11
UTD^G	122	339	141	70	28	10	7	5	0	3	6	33
UTD^P	106	278	132	77	32	16	10	14	3	5	8	91
DFKI^P	386	573	261	151	21	9	4	2	0	1	3	11
Distribution of correctly predicted links over sentence distances												
UTD^G	11	126	39	13	1	0	0	0	0	0	0	0
UTD^P	28	64	23	10	4	2	0	0	0	1	0	0
DFKI^P	0	78	23	18	0	0	0	0	0	0	0	0
Distribution of incorrectly predicted links over sentence distances												
UTD^G	111	213	102	57	27	10	7	5	0	3	6	33
UTD^P	78	214	109	67	28	14	10	14	3	4	8	91
DFKI^P	386	495	238	133	21	9	4	2	0	1	3	11

Table 21: Discourse deixis resolution: distribution of gold/predicted links over the sentence distances between the anaphor and the antecedents.

While Table 20a focuses on the extraction of gold pairs, Table 20b focuses on the extraction of wrong pairs (i.e., wrong links). Specifically, we take the union of the set of wrong pairs extracted by the three resolvers and compute statistics based on the resulting set, which we refer to as S. The columns in Table 20b can be interpreted in the same way as those in Table 20a. For instance, UTD^G shows the percentage of pairs in S extracted by UTD^G, and UTD^G-only shows the percentage of pairs in S that are extracted by UTD^G but not the other two systems.

As can be seen in Table 20b, 1330 erroneous links are established by the three systems, of which 19.5% are established by UTD^G, 20.2% by UTD^P, and 69.0% by DFKI^P; in addition, 13.7% of these erroneous links are only established by UTD^G, 15.6% are only established by UTD^P, and 63.9% are only established by DFKI. These results again reveal that DFKI^P establishes a lot more erroneous links than the UTD systems in each category of anaphors it handles. Interestingly, it attempts to resolve "the", which should not have appeared in the training data as an anaphor. Considering the "x-only" results, we see that there is a fairly large percentage of links in each category that are identified by only one of the three systems, suggesting that the three systems are quite different from each other in their resolution behavior.

To get a better idea of how far a discourse deixis can be from its antecedent, we show in Table 21 the relevant statistics collected from the four test sets. Specifically, the row labeled "Gold" shows the distribution of gold links over the number of sentences between a deixis and its antecedent. (If the sentence distance is 0, it means that the deixis refers to the sentence in which it appears.) As can be seen, the results are consistent with our intuition: a deictic expression most likely has the immediately preceding sentence (i.e., distance = 1) as its referent; in addition, the number of links drops as distance increases. More than 90% of the antecedents are at most four sentences away from their anaphors. In other words, if a discourse deixis resolver simply employs the closest five sentences preceding an anaphor as its candidate antecedents, they should cover more than 90% of the correct antecedents.

The next three rows of Table 21 show the distributions of the links identified by the three resolvers, UTD^G, UTD^P, and DFKI^P. Interestingly, these three distributions all have similar shapes to the gold distribution: they all peak at distance = 1 and generally drop as the sentence distance increases. Next, we tease apart the correct links from the wrong links. The distributions of the correctly predicted links as well as the distributions of the incorrectly predicted links created by the three resolvers over sentence distances are shown in the next two blocks of results. Comparing UTD^P and DFKI^P, we see that DFKI^P does not posit any links between an anaphor and the sentence in which it appears, but it establishes more correct links than UTD^P when the sentence distance is between 1 and 3. Nevertheless, as discussed before, it posits a substantially larger number of erroneous links than the two UTD resolvers.

4.3 Discussion

At first glance it appears that discourse deixis resolution is more challenging the entity coreference resolution and bridging resolution. The reason is that while various string-matching facilities can be used to identify some of the entity coreference relations and bridging relations, they cannot be applied to resolve deictic expressions as there is no content word overlap between a deixis and its antecedent. However, this task has certain characteristics that make it somewhat easier that it seems. First, for antecedents, the unit of annotation is a sentence/utterance; moreover, an antecedent cannot be bigger than a turn (i.e., the utterances produced by a speaker within a turn). These constraints on antecedent annotation can be exploited to significantly reduce the search space of candidate antecedents. Better still, as discussed in the previous subsection, there is a recency constraint that can be exploited to further reduce this search space: a deixis's antecedent typically appears close to it. In contrast, any multi-word expression can be a valid antecedent for an anaphor in entity coreference and bridging; moreover, in some of the entity coreference relations and bridging relations, the two mentions involved can be far apart from each other. Hence, to achieve good performance, a coreference/bridging resolver typically needs to work with a larger search space than a discourse deixis resolver.

The key factor that appears to be limiting the performance of the participating systems is anaphor recognition. The most frequent deictic expressions such as "that", "this", and "it" can also serve as an identity or bridging anaphor, and determining whether a mention is deictic is a key challenge in discourse deixis resolution. As discussed earlier, the low recognition and resolution results achieved by DFKI's system can largely be attributed to its being weak at determining whether a given expression such as "this" or "that" is used deictically and its being overly liberal in classifying these words as deictic and resolving them.

5 Concluding Remarks

We presented a cross-team analysis of the systems that participated in each of the three tracks of the CODI-CRAC 2021 shared task. As noted in the introduction, conducting a systematic analysis that can provide insightful observations is by no means a trivial task. We believe that future cross-team analyses can be improved in a number of ways.

First, any analysis should be based on the *links* identified by a system rather than the output *clusters* they generated. The reason is that a cluster contains both the links identified by a system and the links automatically created via transitive closure. Hence, to better understand the mistakes made by a resolver, we should request the teams to provide the links their systems identify in addition to the clusters they produce so that we can conduct cross-team analyses on the links instead. We do note, however, that this may not be easy for entity-based systems, where a link is established between an anaphor and one of its preceding clusters. For these systems, assumptions may need to be made. For example, when merging takes place, we may assume that a link is established between the anaphor and the member of the cluster that is closest to the anaphor.

Second, to facilitate cross-team comparisons, the teams should be asked to run diagnostic tests provided by the organizers on their systems so that additional insights into their behavior can be gained. For instance, since mention extraction performance has a significant correlation with resolution performance, we will not be able to quantify its impact on resolution performance or directly compare different models in terms of their *resolution* performance that is not affected by their mention detection performance unless we provide a system with *gold* mentions. Hence, a useful diagnostic test, which was employed in the CoNLL-2012 shared task on Unrestricted Multilingual Coreference (Pradhan et al., 2012), involves running the systems on the test data when gold mentions are given. Another useful diagnostic test, which involves running the bridging resolvers and the discourse deixis resolvers on the test data when gold anaphors are given, would allow us to directly compare the resolvers in terms of their antecedent selection performance. While this shared task has a Gold phase for the bridging track and the discourse deixis track in which gold mentions are given, these gold mentions are somewhat different from what one would expect. Specifically, while the participants expected to be given task-specific gold mentions, they were instead provided with gold mentions that were created by taking the union of the three sets of gold mentions collected from the three tracks. Worse still, the gold mentions provided in the discourse deixis track did not even include the antecedents.

Unfortunately, this somewhat unconventional definition of gold mentions was not clearly communicated to the participants and has caused some confusion among them.

Third, our analysis could be improved with an analysis of annotation quality. Strictly speaking, an analysis of annotation quality should not be part of a cross-team analysis, but for this shared task annotation quality may have played a role in the performance of the participating systems given that there are linking decisions that we do not agree with based on our casual inspection of the annotated data. Having said that, we are not sure whether it is possible for us to assess annotation quality since the annotation guidelines are not made available to the participants.

Fourth, from an analysis point of view, it may be a good idea to include LEA (Moosavi and Strube, 2016) as one of the evaluation metrics. As Moosavi and Strube point out, it is not easy to interpret the scores provided by existing scorers such as MUC, B^3, and $CEAF_e$ and LEA is designed to partially address this problem.

Fifth, our analysis did not focus on the differences among the four datasets. For instance, in discourse deixis resolution, UTD^G achieved significantly better results than UTD^P on Persuasion but not the other datasets. A cross-dataset analysis could shed light on why systems exhibit different trends on different datasets. Having said that, the organizers should take an active role in explaining to the participants the differences among the different datasets and the unique challenges associated with each of them so that the participants know why these four datasets were chosen, rather than have them figure these differences out on their own.

Finally, to facilitate cross-team analyses, the organizers should include as much relevant information in the system prediction files that they make available to the participants as possible. Currently, these files merely contain the pairwise predictions made by the systems as well as the gold links they missed. Some potentially useful information that can also be provided in these files includes the sentence/turn distance between the mention pairs and their surrounding contexts. While this information can be extracted by the participants from the raw system outputs, simply laying the burden on them will likely deter them from conducting an analysis.

Acknowledgments

This work was supported in part by NSF Grant IIS-1528037. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views or official policies, either expressed or implied, of the NSF.

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*, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the First International Conference on Language Resources and Evaluation Workshop on Linguistic Coreference*, pages 563–566.

Yufang Hou. 2018. Enhanced word representations for bridging anaphora resolution. 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 1–7, New Orleans, Louisiana. Association for Computational Linguistics.

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.

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*, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hongjin Kim, Damrin Kim, and Harksoo Kim. 2021. The pipeline model for resolution of anaphoric reference and resolution of entity reference. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, 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*, Punta Cana, Dominican Republic. 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.

Jonathan K. Kummerfeld and Dan Klein. 2013. Error-driven analysis of challenges in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 265–277, Seattle, Washington, USA. Association for Computational Linguistics.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Jing Lu and Vincent Ng. 2020. Conundrums in entity coreference resolution: Making sense of the state of the art. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6620–6631, Online. Association for Computational Linguistics.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Berlin, Germany. Association for Computational Linguistics.

Onkar Pandit, Pascal Denis, and Liva Ralaivola. 2020. Integrating knowledge graph embeddings to improve mention representation for bridging anaphora resolution. In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 55–67, Barcelona, Spain (online). 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 *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon, USA. Association for Computational Linguistics.

Joseph Renner, Priyansh Trivedi, Gaurav Maheshwari, Rémi Gilleron, and Pascal Denis. 2021. An end-to-end approach for full bridging resolution. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

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, Online. Association for Computational Linguistics.

Liyan Xu and Jinho D. Choi. 2021. Adapted end-to-end coreference resolution system for anaphoric identities in dialogues. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, Punta Cana, Dominican Republic. Association for Computational Linguistics.