End-to-End Neural Discourse Deixis Resolution in Dialogue

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

Abstract

We adapt Lee et al.'s (2018) span-based entity coreference model to the task of end-to-end discourse deixis resolution in dialogue, specifically by proposing extensions to their model that exploit task-specific characteristics. The resulting model, `dd-utt`, achieves state-of-the-art results on the four datasets in the CODI-CRAC 2021 shared task.

1 Introduction

Discourse deixis (DD) resolution, also known as abstract anaphora resolution, is an under-investigated task that involves resolving a deictic anaphor to its antecedent. A *deixis* is a reference to a discourse entity such as a proposition, a description, an event, or a speech act (Webber, 1991). DD resolution is arguably more challenging than the extensively-investigated entity coreference resolution task. Recall that in entity coreference, the goal is to cluster the entity mentions in narrative text or dialogue, which are composed of pronouns, names, and nominals, so that the mentions in each cluster refer to the same real-world entity. Lexical overlap is a strong indicator of entity coreference, both among names (e.g., "President Biden", "Joe Biden") and in the resolution of nominals (e.g., linking "the president" to "President Biden"). DD resolution, on the other hand, can be viewed as a generalized case of event coreference involving the clustering of deictic anaphors, which can be pronouns or nominals, and clauses, such that the mentions in each cluster refer to the same real-world proposition/event/speech act, etc. The first example in Figure 1 is an example of DD resolution in which the deictic anaphor "the move" refers to Salomon's act of issuing warrants on shares described in the preceding sentence. DD resolution is potentially more challenging than entity coreference resolution because (1) DD resolution involves understanding *clause* semantics since antecedents are clauses, and clause semantics are arguably harder to encode than noun phrase semantics; and (2) string matching plays little role in DD resolution, unlike in entity coreference.

Salomon Brothers International Ltd. announced it will issue warrants on shares of Hong Kong Telecommunications Ltd. *The move* closely follows a similar offer by Salomon of warrants for shares of Hongkong & Shanghai Banking Corp.

A: Would you donate to Save the Children?
B: **Yes, I will do $10 to both.**
B: I am of a tight budget, but I do make room for good causes.
A: Thank you very much.
A: The children will appreciate *it*.

Figure 1: Examples of discourse deixis resolution. In each example, the deictic anaphor is italicized and the antecedent is boldfaced.

In this paper, we focus on end-to-end DD resolution in dialogue. The second example in Figure 1 shows a dialogue between A and B in which the deictic anaphor "it" refers to the utterance by B in which s/he said s/he would donate $10. While the deictic anaphors in dialogue are also composed of pronouns and nominals, the proportion of pronominal deictic anaphors in dialogue is much higher than that in narrative text. For instance, while 76% of the deictic anaphors in two text corpora (ARRAU RST and GNOME) are pronominal, the corresponding percentage rises to 93% when estimated based on seven dialogue corpora (TRAINS91, TRAINS93, PEAR, and the four CODI-CRAC 2021 development sets). In fact, the three pronouns "that", "this", and "it" alone comprise 89% of the deictic anaphors in these dialogue corpora. The higher proportion of pronominal deictic anaphors potentially makes DD resolution in dialogue more challenging than those in text: since a pronoun is semantically empty, the successful resolution of a pronominal deictic anaphor depends entirely on proper understanding of its context. In addition, it also makes DD *recognition* more challenging in dialogue. For instance, while the head of a non-pronominal phrase can often be exploited to determine whether it is a deictic anaphor (e.g., "the

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 11322 - 11334
December 7-11, 2022 ©2022 Association for Computational Linguistics

man" cannot be a deictic anaphor, but "the move" can), such cues are absent in pronouns.

Since DD resolution can be cast as a generalized case of event coreference, a natural question is: how successful would a state-of-the-art entity coreference model be when applied to DD resolution? Recently, Kobayashi et al. (2021) have applied Xu and Choi's (2020) re-implementation of Lee et al.'s span-based entity coreference model to resolve the deictic anaphors in the DD track of the CODI-CRAC 2021 shared task after augmenting it with a *type prediction* model (see Section 4). Not only did they achieve the highest score on each dataset, but they beat the second-best system (Anikina et al., 2021), which is a non-span-based neural approach combined with hand-crafted rules, by a large margin. These results suggest that a span-based approach to DD resolution holds promise.

Our contributions in this paper are three-fold. First, we investigate whether *task-specific* observations can be exploited to extend a span-based model originally developed for entity coreference to improve its performance for end-to-end DD resolution in dialogue. Second, our extensions are effective in improving model performance, allowing our model to achieve state-of-the-art results on the CODI-CRAC 2021 shared task datasets. Finally, we present an empirical analysis of our model, which, to our knowledge, is the first analysis of a state-of-the-art span-based DD resolver.

2 Related Work

Broadly, existing approaches to DD resolution can be divided into three categories, as described below.

Rule-based approaches. Early systems that resolve deictic expressions are rule-based (Eckert and Strube, 2000; Byron, 2002; Navarretta, 2000). Specifically, they use predefined rules to extract anaphoric mentions, and select antecedent for each extracted anaphor based on the dialogue act types of each candidate antecedent.

Non-neural learning-based approaches. Early non-neural learning-based approaches to DD resolution use hand-crafted feature vectors to represent mentions (Strube and Müller, 2003; Müller, 2008). A classifier is then trained to determine whether a pair of mentions is a valid antecedent-anaphor pair.

Deep learning-based approaches. Deep learning has been applied to DD resolution. For instance, Marasović et al. (2017) and Anikina et al. (2021) use a Siamese neural network, which takes as input the embeddings of two sentences, one containing a deictic anaphor and the other a candidate antecedent, to score each candidate antecedent and subsequently rank the candidate antecedents based on these scores. In addition, motivated by the recent successes of Transformer-based approaches to entity coreference (e.g., Kantor and Globerson (2019)), Kobayashi et al. (2021) have recently proposed a Transformer-based approach to DD resolution, which is an end-to-end coreference system based on SpanBERT (Joshi et al., 2019, 2020). Their model jointly learns mention extraction and DD resolution and has achieved state-of-the-art results in the CODI-CRAC 2021 shared task.

3 Corpora

We use the DD-annotated corpora provided as part of the CODI-CRAC 2021 shared task. For training, we use the official training corpus from the shared task (Khosla et al., 2021), ARRAU (Poesio and Artstein, 2008), which consists of three conversational sub-corpora (TRAINS-93, TRAINS-91, PEAR) and two non-dialogue sub-corpora (GNOME, RST). For validation and evaluation, we use the official development sets and test sets from the shared task. The shared task corpus is composed of four well-known conversational datasets: AMI (McCowan et al., 2005), LIGHT (Urbanek et al., 2019), Persuasion (Wang et al., 2019), and Switchboard (Godfrey et al., 1992). Statistics on these corpora are provided in Table 1.

4 Baseline Systems

We employ three baseline systems.

The first baseline, `coref-hoi`, is Xu and Choi's (2020) re-implementation of Lee et al.'s (2018) widely-used end-to-end entity coreference model. The model ranks all text spans of up to a predefined length based on how likely they correspond to entity mentions. For each top-ranked span x, the model learns a distribution $P(y)$ over its antecedents $y \in \mathcal{Y}(x)$, where $\mathcal{Y}(x)$ includes a dummy antecedent ϵ and every preceding span:

$$P(y) = \frac{e^{s(x,y)}}{\sum_{y' \in \mathcal{Y}(x)} e^{s(x,y')}}$$

where $s(x, y)$ is a pairwise score that incorporates two types of scores: (1) $s_m(\cdot)$, which indicates how likely a span is a mention, and (2) $s_c(\cdot)$ and $s_a(\cdot)$, which indicate how likely two spans refer

		Total #docs	Total #sents	Total #turns	Avg. #sents	Avg. #toks per sent	Avg. #turns	Avg. #ana	Avg. #ante	Avg. #speakers per doc
ARRAU	train	552	22406	-	40.6	15.5	-	2.9	4.8	-
LIGHT	dev	20	908	280	45.4	12.7	14.0	3.1	4.2	2.0
	test	21	923	294	44.0	12.8	14.0	3.8	4.6	2.0
AMI	dev	7	4139	2828	591.3	8.2	404.0	32.9	42.0	4.0
	test	3	1967	1463	655.7	9.3	487.7	39.3	47.3	4.0
Pers.	dev	21	812	431	38.7	11.3	20.5	4.5	4.5	2.0
	test	28	1139	569	40.7	11.1	20.3	4.4	4.8	2.0
Swbd.	dev	11	1342	715	122.0	11.2	65.0	11.5	15.9	2.0
	test	22	3652	1996	166.0	9.6	90.7	12.0	14.7	2.0

Table 1: Statistics on the datasets.

to the same entity[1] ($s_a(x, \epsilon) = 0$ for dummy antecedents):

$$\begin{aligned} s(x, y) &= s_m(x) + s_m(y) + s_c(x, y) + s_a(x, y) && (1) \\ s_m(x) &= \texttt{FFNN}_m(g_x) && (2) \\ s_c(x, y) &= g_x^\top W_c g_y && (3) \\ s_a(x, y) &= \texttt{FFNN}_c(g_x, g_y, g_x \circ g_y, \phi(x, y)) && (4) \end{aligned}$$

where g_x and g_y are the vector representations of x and y, W_c is a learned weight matrix for bilinear scoring, FFNN $(\cdot)$ is a feedforward neural network, and $\phi(\cdot)$ encodes features. Two features are used, one encoding speaker information and the other the segment distance between two spans.

The second baseline, UTD_NLP[2], is the top-performing system in the DD track of the CODI-CRAC 2021 shared task (Kobayashi et al., 2021). It extends coref-hoi with a set of modifications. Two of the most important modifications are: (1) the addition of a sentence distance feature to $\phi(\cdot)$, and (2) the incorporation into coref-hoi a *type prediction* model, which predicts the type of a span. The possible types of a span i are: ANTECEDENT (if i corresponds to an antecedent), ANAPHOR (if i corresponds to an anaphor), and NULL (if it is neither an antecedent nor an anaphor). The types predicted by the model are then used by coref-hoi as follows: only spans predicted as ANAPHOR can be resolved, and they can only be resolved to spans predicted as ANTECEDENT. Details of how the type prediction model is trained can be found in Kobayashi et al. (2021).

The third baseline, coref-hoi-utt, is essentially the first baseline except that we restrict the candidate antecedents to be *utterances*. This restriction is motivated by the observation that the antecedents of the deictic anaphors in the CODI-CRAC 2021 datasets are all utterances. To see what an utterance is, consider again the second example in Figure 1. Each line in this dialogue is an utterance. As can be seen, an utterance roughly corresponds to a sentence, although it can also be a text fragment or simply an interjection (e.g., “uhhh”). While by definition the antecedent of a deictic anaphor can be any clause, the human annotators of the DD track of the CODI-CRAC 2021 shared task decided to restrict the unit of annotation to utterances because (1) based on previous experience it was difficult to achieve high inter-annotator agreement when clauses are used as the annotation unit (Poesio and Artstein, 2008); and (2) unlike the sentences in narrative text, which can be composed of multiple clauses and therefore can be long, the utterances in these datasets are relatively short and can reliably be used as annotation units. From a modeling perspective, restricting candidate antecedents also has advantages. First, it substantially reduces the number of candidate antecedents and therefore memory usage, allowing our full model to fit into memory. Second, it allows us to focus on *resolution* rather than *recognition* of deictic anaphors, as the recognition of clausal antecedents remains a challenging task, especially since existing datasets for DD resolution are relatively small compared to those available for entity coreference (e.g., OntoNotes (Hovy et al., 2006)).

5 Model

Next, we describe our resolver, dd-utt, which augments coref-hoi-utt with 10 extensions.

E1. Modeling recency. Unlike in entity coreference, where two coreferent names (e.g., “Joe Biden”, “President Biden”) can be far apart from

[1] See Lee et al. (2018) for a description of the differences between $s_c(\cdot)$ and $s_a(\cdot)$,

[2] For an analysis of this and other resolvers competing in the CODI-CRAC 2021 shared task, see Li et al. (2021).

each other in the corresponding document (because names are non-anaphoric), the distance between a deictic anaphor and its antecedent is comparatively smaller. To model recency, we restrict the set of candidate antecedents of an anaphor to be the utterance containing the anaphor as well as the preceding 10 utterances, the choice of which is based on our observation of the development data, where the 10 closest utterances already cover 96–99% of the antecedent-anaphor pairs.

E2. Modeling distance. While the previous extension allows us to restrict our attention to candidate antecedents that are close to the anaphor, it does not model the fact that the likelihood of being the correct antecedent tends to increase as its distance from the anaphor decreases. To model this relationship, we subtract the term $\gamma_1 Dist(x, y)$ from $s(x, y)$ (see Equation (1)), where $Dist(x, y)$ is the utterance distance between anaphor x and candidate antecedent y and γ_1 is a tunable parameter that controls the importance of utterance distance in the resolution process. Since $s(x, y)$ is used to rank candidate antecedents, modeling utterance distance by updating $s(x, y)$ will allow distance to have a direct impact on DD resolution.

E3. Modeling candidate antecedent length. Some utterances are pragmatic in nature and do not convey any important information. Therefore, they cannot serve as antecedents of deictic anaphors. Examples include "Umm", "Ahhhh... okay", "that's right", and "I agree". Ideally, the model can identify such utterances and prevent them from being selected as antecedents. We hypothesize that we could help the model by modeling such utterances. To do so, we observe that such utterances tend to be short and model them by penalizing shorter utterances. Specifically, we subtract the term $\gamma_2 \frac{1}{Length(y)}$ from $s(x, y)$, where $Length(y)$ is the number of words in candidate antecedent y and γ_2 is a tunable parameter that controls the importance of candidate antecedent length in resolution.

E4. Extracting candidate anaphors. As mentioned before, the deictic anaphors in dialogue are largely composed of pronouns. Specifically, in our development sets, the three pronouns "that", "this", and 'it' alone account for 74–88% of the anaphors. Consequently, we extract candidate deictic anaphors as follows: instead of allowing each span of length n or less to be a candidate anaphor, we only allow a span to be a candidate anaphor if its underlying word/phrase has appeared at least once in the training set as a deictic anaphor.

E5. Predicting anaphors. Now that we have the candidate anaphors, our next extension involves predicting which of them are indeed deictic anaphors. To do so, we retrain the type prediction model in `UTD_NLP`, which is a FFNN that takes as input the (contextualized) span representation g_i of candidate anaphor i and outputs a vector ot_i of dimension 2 in which the first element denotes the likelihood that i is a deictic anaphor and the second element denotes the likelihood that i is not a deictic anaphor. i is predicted as a deictic anaphor if and only if the value of the first element of ot_i is bigger than its second value:

$$\begin{aligned} ot_i &= \texttt{FFNN}(g_i) \\ t_i &= \arg\max_{x \in \{\text{A,NA}\}} ot_i(x) \end{aligned}$$

where A (ANAPHOR) and NA (NON-ANAPHOR) are the two possible types. Following `UTD_NLP`, this type prediction model is jointly trained with the resolution model. Specifically, we compute the cross-entropy loss using ot_i, multiply it by a type loss coefficient λ, and add it to the loss function of `coref-hoi-utt`. λ is a tunable parameter that controls the importance of type prediction relative to DD resolution.

E6. Modeling the relationship between anaphor recognition and resolution. In principle, the model should resolve a candidate anaphor to a non-dummy candidate antecedent if it is predicted to be a deictic anaphor by the type prediction model. However, type prediction is not perfect, and enforcing this consistency constraint, which we will refer to as **C1**, will allow errors in type prediction to be propagated to DD resolution. For example, if a non-deictic anaphor is misclassified by the type prediction model, then it will be (incorrectly) resolved to a non-dummy antecedent. To alleviate error propagation, we instead enforce **C1** in a *soft* manner. To do so, we define a penalty function p_1, which imposes a penalty on span i if **C1** is violated (i.e., a deictic anaphor is resolved to the dummy antecedent), as shown below:

$$p_1(i) = \begin{cases} 0 & \arg\max_{y_{is} \in \mathcal{Y}} t_i = \text{NA} \\ ot_i(\text{A}) - ot_i(\text{NA}) & \text{otherwise} \end{cases}$$

Intuitively, p_1 estimates the minimum amount to be adjusted so that span i's type is not ANAPHOR.

We incorporate p_1 into the model as a penalty term in s (Equation (1)). Specifically, we redefine $s(i,j)$ when $j=\epsilon$, as shown below:

$$s(i,\epsilon) = s(i,\epsilon) - [\gamma_3 p_1(i)]$$

where γ_3 is a positive constant that controls the hardness of **C1**. The smaller γ_3 is, the softer **C1** is. Intuitively, if **C1** is violated, $s(i,\epsilon)$ will be lowered by the penalty term, and the dummy antecedent will less likely be selected as the antecedent of i.

E7. Modeling the relationship between non-anaphor recognition and resolution. Another consistency constraint that should be enforced is that the model should resolve a candidate anaphor to the dummy antecedent if it is predicted as a non-deictic anaphor by the type prediction model. As in Extension E6, we will enforce this constraint, which we will refer to as **C2**, in a soft manner by defining a penalty function p_2, as shown below:

$$p_2(i) = \begin{cases} ot_i(\text{NA}) - ot_i(\text{A}) & \underset{y_{is}\in\mathcal{Y}}{\arg\max}\, t_i = \text{NA} \\ 0 & \text{otherwise} \end{cases}$$

Then we redefine $s(i,j)$ when $j \neq \epsilon$ as follows:

$$s(i,j) = s(i,j) - [\gamma_4 p_1(i)]$$

where γ_4 is a positive constant that controls the hardness of **C2**. Intuitively, if **C2** is violated, $s(i,j)$ will be lowered by the penalty term, and j will less likely be selected as the antecedent of i.

E8. Encoding candidate anaphor context. Examining Equation (1), we see that $s(x,y)$ is computed based on the span representations of x and y. While these span representations are contextualized, the contextual information they encode is arguably limited. As noted before, most of the deictic anaphors in dialogue are pronouns, which are semantically empty. As a result, we hypothesize that we could improve the resolution of these deictic anaphors if we explicitly modeled their contexts. Specifically, we represent the context of a candidate anaphor using the embedding of the utterance in which it appears and add the resulting embedding as features to both the bilinear score $s_c(x,y)$ and the concatenation-based score $s_a(x,y)$:

$$s_c(x,y) = g_x^\top W_c g_y + g_s^\top W_s g_y$$
$$s_a(x,y) = \texttt{FFNN}_c(g_x, g_y, g_x \circ g_y, g_s, \phi(x,y))$$

where W_c and W_s are learned weight matrices, g_s is the embedding of the utterance s in which candidate anaphor x appears, and $\phi(x,y)$ encodes the relationship between x and y as features.

Filling words
yeah, okay, ok, uh, right, so, hmm, well, um, oh, mm, yep, hi, ah, whoops, alright, shhhh, yes, ay, hello, aww, alas, ye, aye, uh-huh, huh, wow, www, no, and, but, again, wonderful, exactly, absolutely, actually, sure thanks, awesome, gosh, ooops
Reporting verbs
command, mention, demand, request, reveal, believe, guarantee, guess, insist, complain, doubt, estimate, warn, learn, realise, persuade, propose, announce, advise, imagine, boast, suggest, remember, claim, describe, see, understand, discover, answer, wonder, recommend, beg, prefer, suppose, comment, think, argue, consider, swear, ask, agree, explain, report, know, tell, decide, discuss, repeat, invite, reply, expect, forget, add, fear, hope, say, feel, observe, remark, confirm, threaten, teach, forbid, admit, promise, deny, state, mean, instruct

Table 2: Lists of filtered words.

E9. Encoding the relationship between candidate anaphors and antecedents. As noted in Extension E8, $\phi(x,y)$ encodes the relationship between candidate anaphor x and candidate antecedent y. In `UTD_NLP`, $\phi(x,y)$ is composed of three features, including two features from `coref-hoi-utt` (i.e., the speaker id and the *segment* distance between x and y) and one feature that encodes the *utterance* distance between them. Similar to the previous extension, we hypothesize that we could better encode the relationship between x and y using additional features. Specifically, we incorporate an additional feature into $\phi(x,y)$ that encodes the utterance distance between x and y. Unlike the one used in `UTD_NLP`, this feature aims to more accurately capture proximity by ignoring *unimportant* sentences (i.e., those that contain only interjections, filling words, reporting verbs, and punctuation) when computing utterance distance. The complete list of filling words and reporting verbs that we filter can be found in Table 2.

E10. Encoding candidate antecedents. In `coref-hoi-utt`, a candidate antecedent is simply encoded using its span representation. We hypothesize that we could better encode a candidate antecedent using additional *features*. Specifically, we employ seven features to encode a candidate antecedent y and incorporate them into $\phi(x,y)$: (1) the number of words in y; (2) the number of nouns in y; (3) the number of verbs in y; (4) the number of adjectives in y; (5) the number of content word overlaps between y and the portion of the utterance containing the anaphor that precedes the anaphor; (6) whether y is the longest among the candidate antecedents; and (7) whether y has the largest number of content word overlap (as computed in Feature 5) among the candidate antecedents. Like Extension E3, some features implicitly encode the length of a candidate antecedent. Despite this redundancy, we believe the redundant information could be exploited by the model differently and may therefore have varying degrees of impact on it.

6 Evaluation

6.1 Experimental Setup

Evaluation metrics. We obtain the results of DD resolution using the Universal Anaphora Scorer (Yu et al., 2022b). Since DD resolution is viewed as a generalized case of event coreference, the scorer reports performance in terms of CoNLL score, which is the unweighted average of the F-scores of three coreference scoring metrics, namely MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), and CEAF$_e$ (Luo, 2005). In addition, we report the results of deictic anaphor recognition. We express recognition results in terms of Precision (P), Recall (R) and F-score, considering an anaphor correctly recognized if it has an exact match with a gold anaphor in terms of boundary.

Model training and parameter tuning. For `coref-hoi` and `coref-hoi-utt`, we use SpanBERT$_{\text{Large}}$ as the encoder and reuse the hyperparameters from Xu and Choi (2020) with the only exception of the maximum span width: for `coref-hoi`, we increase the maximum span width from 30 to 45 in order to cover more than 97% of the antecedent spans; for `coref-hoi-utt` we use 15 as the maximum span width, which covers more than 99% of the anaphor spans in the training sets. For `UTD_NLP`, we simply take the outputs produced by the model on the test sets and report the results obtained by running the scorer on the outputs.[3] For `dd-utt`, we use SpanBERT$_{\text{Large}}$ as the encoder. Since we do not rely on span enumerate to generate candidate spans, the maximum span width can be set to any arbitrary number that is large enough to cover all candidate antecedents and anaphors. In our case, we use 300 as our maximum span width. We tune the parameters (i.e., λ, γ_1, γ_2, γ_3, γ_4) using grid search to maximize CoNLL score on development data. For the type loss coefficient, we search out of {0.2, 0.5, 1, 200, 500, 800, 1200, 1600}, and for γ, we search out of {1, 5, 10}. We reuse other hyperparameters from Xu and Choi (2020).

All models are trained for 30 epochs with a dropout rate of 0.3 and early stopping. We use 1×10^{-5} as our BERT learning rate and 3×10^{-4} as our task learning rate. Each experiment is run using a random seed of 11 and takes less than three hours to train on an NVIDIA RTX A6000 48GB.

Train-dev partition. Since we have four test sets, we use ARRAU and all dev sets other than the one to be evaluated on for model training and the remaining dev set for parameter tuning. For example, when evaluating on AMI$_{\text{test}}$, we train models on ARRAU, LIGHT$_{\text{dev}}$, Persuasion$_{\text{dev}}$ and Switchboard$_{\text{dev}}$ and use AMI$_{\text{dev}}$ for tuning.

6.2 Results

Recall that our goal is to perform end-to-end DD resolution, which corresponds to the Predicted evaluation setting in the CODI-CRAC shared task.

Overall performance. Recognition results (expressed in F-score) and resolution results (expressed in CoNLL score) of the three baselines and our model on the four test sets are shown in Table 3, where the Avg. columns report the macro-averages of the corresponding results on the four test sets, and the parameter settings that enable our model to achieve the highest CoNLL scores on the development sets are shown in Table 4. Since `coref-hoi` and `coref-hoi-utt` do not explicitly identify deictic anaphors, we assume that all but the first mentions in each output cluster are anaphors when computing recognition precision; and while `UTD_NLP` (the top-performing system in the shared task) does recognize anaphors, we still make the same assumption when computing its recognition precision since the anaphors are not explicitly marked in the output (recall that we computed results of `UTD_NLP` based on its outputs).

We test the statistical significance among the four models using two-tailed Approximate Randomization (Noreen, 1989). For recognition, the models are statistically indistinguishable from each other w.r.t. their Avg. score ($p < 0.05$). For resolution, `dd-utt` is highly significantly better than the baselines w.r.t. Avg. ($p < 0.001$), while the three baselines are statistically indistinguishable from each other. These results suggest that (1) `dd-utt`'s superior resolution performance stems from better antecedent selection, not better anaphor

[3] Since the shared task participants were allowed to submit their system outputs multiple times to the server to obtain results on the test sets, `UTD_NLP`'s results could be viewed as results obtained by tuning parameters on the test sets.

	Resolution					Recognition				
	LIGHT	AMI	Pers.	Swbd.	Avg.	LIGHT	AMI	Pers.	Swbd.	Avg.
UTD_NLP	42.7	35.4	39.6	35.4	38.3	70.1	61.0	69.9	68.1	67.3
coref-hoi	42.7	30.7	49.7	35.4	39.6	70.9	49.3	67.8	61.9	62.5
coref-hoi-utt	42.3	35.0	53.3	34.1	41.2	70.3	52.4	71.0	60.6	63.6
dd-utt	48.2	43.5	54.9	47.2	48.5	71.3	56.9	71.4	65.2	66.2

Table 3: Resolution and recognition results on the four test sets.

	LIGHT	AMI	Pers.	Swbd.
Type loss coefficient λ	800	800	800	800
γ_1	1	1	1	1
γ_2	1	1	1	1
γ_3	5	10	10	5
γ_4	5	5	5	5

Table 4: Parameter values enabling dd-utt to achieve the best CoNLL score on each development set.

recognition; and (2) the restriction of candidate antecedents to utterances in coref-hoi-utt does not enable the resolver to yield significantly better resolution results than coref-hoi.

Per-anaphor results. Next, we show the recognition and resolution results of the four models on the most frequently occurring deictic anaphors in Table 5 after micro-averaging them over the four test sets. Not surprisingly, "that" is the most frequent deictic anaphor on the test sets, appearing as an anaphor 402 times on the test sets and contributing to 68.8% of the anaphors. This is followed by "it" (16.3%) and "this" (4.3%). Only 8.9% of the anaphors are not among the top four anaphors.

Consider first the recognition results. As can be seen, "that" has the highest recognition F-score among the top anaphors. This is perhaps not surprising given the comparatively larger number of "that" examples the models are trained on. While "it" occurs more frequently than "this" as a deictic anaphor, its recognition performance is lower than that of "this". This is not surprising either: "this", when used as a pronoun, is more likely to be deictic than "it", although both of them can serve as a coreference anaphor and a bridging anaphor. In other words, it is comparatively more difficult to determine whether a particular occurrence of "it" is deictic. Overall, UTD_NLP recognizes more anaphors than the other models.

Next, consider the resolution results. To obtain the CoNLL scores for a given anaphor, we retain all and only those clusters containing the anaphor in both the gold partition and the system partition and apply the official scorer to them. Generally, the more frequently occurring an anaphor is, the better its resolution performance is. Interestingly, for the "Others" category, dd-utt achieves the highest resolution results despite having the lowest recognition performance. In contrast, while UTD_NLP achieves the best recognition performance on average, its resolution results are among the worst.

Per-distance results. To better understand how resolution results vary with the utterance distance between a deictic anaphor and its antecedent, we show in Table 6 the number of correct and incorrect links predicted by the four models for each utterance distance on the test sets. For comparison purposes, we show at the top of the table the distribution of gold links over utterance distances. Note that a distance of 0 implies that the anaphor refers to the utterance in which it appears.

A few points deserve mention. First, the distribution of gold links is consistent with our intuition: a deictic anaphor most likely has the immediately preceding utterance (i.e., distance = 1) as its referent. In addition, the number of links drops as distance increases, and more than 90% of the antecedents are at most four utterances away from their anaphors. Second, although UTD_NLP recognizes more anaphors than the other models, it is the most conservative w.r.t. link identification, predicting the smallest number of correct and incorrect links for almost all of the utterance distances. Third, dd-utt is better than the other models at (1) identifying short-distance anaphoric dependencies, particularly when distance ≤ 1, and (2) positing fewer erroneous long-distance anaphoric dependencies. These results provide suggestive evidence of dd-utt's success at modeling recency and distance explicitly. Finally, these results suggest that resolution difficulty increases with distance: except for UTD_NLP, none of the models can successfully recognize a link when distance > 5.

Ablation results. To evaluate the contribution of each extension presented in Section 5 to dd-utt's resolution performance, we show in Table 7 ablation results, which we obtain by removing one

		Resolution				Recognition			
anaphor	count	UTD_NLP	coref-hoi	coref-hoi-utt	dd-utt	UTD_NLP	coref-hoi	coref-hoi-utt	dd-utt
that	402	48.7	49.1	50.4	60.7	79.1	72.7	72.9	76.8
it	95	21.8	27.5	28.5	37.0	36.5	36.4	37.0	35.5
this	25	20.7	25.1	26.9	25.8	51.7	49.2	49.3	56.5
which	10	4.5	8.2	14.9	14.4	33.3	28.6	42.1	35.3
Others	52	33.0	39.6	41.2	48.5	33.8	28.1	40.0	10.3

Table 5: Per-anaphor recognition and resolution results on the test sets.

	0	1	2	3	4	5	>5
Distribution of gold links over utterance distances							
Gold	90	209	97	46	21	8	19
Distribution of correctly predicted links							
UTD_NLP	28	64	23	9	4	2	1
coref-hoi	22	108	42	13	5	3	0
coref-hoi-utt	23	118	49	11	6	2	0
dd-utt	40	142	45	17	5	3	0
Distribution of incorrectly predicted links							
UTD_NLP	16	56	31	24	9	6	52
coref-hoi	83	148	55	33	22	13	58
coref-hoi-utt	64	123	66	36	21	10	48
dd-utt	43	131	57	24	6	7	7

Table 6: Distribution of links over the utterance distances between the anaphor and the antecedents.

	LIGHT	AMI	Pers.	Swbd.	Avg.
dd-utt	48.2	43.5	54.9	47.2	48.5
– E1	47.1	36.4	53.6	46.2	45.8
– E2	47.4	40.0	56.5	48.9	48.2
– E3	45.3	44.6	53.7	49.0	48.1
– E4	46.4	42.8	56.7	45.6	47.9
– E5	43.6	43.9	50.2	47.4	46.3
– E6	46.1	43.1	50.8	47.2	46.8
– E7	48.6	43.6	56.0	49.4	49.4
– E8	43.3	39.4	52.5	50.1	46.3
– E9	44.6	43.2	52.1	47.7	46.9
– E10	47.3	39.3	57.5	49.7	48.5

Table 7: Resolution results of ablated models.

extension at a time from `dd-utt` and retraining it. For ease of comparison, we show in the first row of the table the CoNLL scores of `dd-utt`.

A few points deserve mention. First, when E1 (Modeling recency) is ablated, we use as candidate antecedents the 10 highest-scoring candidate antecedents for each candidate anaphor according to $s_c(x, y)$ (Equation (3)). Second, when one of E2, E3, E6, and E7 is ablated, we set the corresponding λ to zero. Third, when E4 is ablated, candidate anaphors are extracted in the same way as in `coref-hoi` and `coref-hoi-utt`, where the top spans learned by the model will serve as candidate anaphors. Fourth, when E5 is ablated, E6 and E7 will also be ablated because the penalty functions p_1 and p_2 need to be computed based on the output of the type prediction model in E5.

We use two-tailed Approximate Randomization to determine which of these ablated models is statistically different from `dd-utt` w.r.t. the Avg. score. Results show that except for the model in which E1 is ablated, all of the ablated models are statistically indistinguishable from `dd-utt` ($p < 0.05$). Note that these results do *not* imply that nine of the extensions fail to contribute positively to `dd-utt`'s resolution performance: it only means that none of them is useful in the presence of other extensions w.r.t. Avg. We speculate that (1) some of these extensions model overlapping phenomena (e.g., both E2 and E9 model utterance distance); (2) when the model is retrained, the learner manages to adjust the network weights so as to make up for the potential loss incurred by ablating an extension; and (3) large fluctuations in performance can be observed on individual datasets in some of the experiments, but they may just disappear after averaging. Experiments are needed to determine the reason.

6.3 Error Analysis

Below we analyze the errors made by `dd-utt`.

DD anaphora recognition precision errors. A common type of recognition precision errors involves misclassifying a coreference anaphor as a deictic anaphor. Consider the first example in Figure 2, in which the pronoun "that" is a coreference anaphor with "voice recognition" as its antecedent but is misclassified as a deictic anaphor with the whole sentence as its antecedent. This type of error occurs because virtually all of the frequently occurring deictic anaphors, including "that", "it", "this", and "which", appear as a coreference anaphor in some contexts and as a deictic anaphor in other contexts, and distinguishing between the two different uses of these anaphors could be challenging.

DD anaphor recognition recall errors. Consider the second example in Figure 2, in which

A: The design should minimize R_S_I and be easy to locate and we were still slightly ambivalent as to whether to use *voice recognition* there, though *that* did seem to be the favored strategy, but there was also, on the sideline, the thought of maybe having a beeper function.

A: **Sounds like a blessed organization.**
B: Yes, *it* does.

A: **Did you know they've won over 7 different awards for their charitable work?**
A: *As a former foster kid, it makes me happy to see this place bring such awareness to the issues and needs of our young.*
B: I am not surprised to hear *that* at all.

Figure 2: Examples illustrating the three majors types of errors made by dd-utt.

"it" is a deictic anaphor that refers to the boldfaced utterance, but dd-utt fails to identify this and many other occurrences of "it" as deictic, probably because "it" is more likely to be a coreference anaphor than a deictic anaphor: in the dev sets, 80% of the occurrences of "it" are coreference anaphors while only 5% are deictic anaphors.

DD resolution precision errors. A major source of DD resolution precision errors can be attributed to the model's failure in properly understanding the context in which a deictic anaphor appears. Consider the third example in Figure 2, in which "that" is a deictic anaphor that refers to the boldfaced utterance. While dd-utt correctly identifies "that" as a deictic anaphor, it erroneously posits the italicized utterance as its antecedent. This example is interesting in that *without* looking at the boldfaced utterance, the italicized utterance is a plausible antecedent for "that" because "I am not surprised to hear that at all" can be used as a response to almost every statement. However, when both the boldfaced utterance and the italicized utterance are taken into consideration, it is clear that the boldfaced utterance is the correct antecedent for "that" because winning over seven awards for some charitable work is certainly more surprising than seeing a place bring awareness to the needs of the young. Correctly resolving this anaphor, however, requires modeling the emotional implication of its context.

6.4 Further Analysis

Next, we analyze the deictic anaphors correctly resolved by dd-utt but erroneously resolved by the baseline resolvers.

The example shown in Figure 3 is one such case. In this example, dd-utt successfully extracts the anaphor "that" and resolves it to the correct antecedent "Losing one decimal place, that is okay". UTD_NLP fails to extract "that" as a deictic anaphor. While coref-hoi correctly extracts the anaphor, it incorrectly selects "You want your rating to be a two?" as the antecedent. From a cursory look at this example one could infer that this candidate antecedent is highly unlikely to be the correct antecedent since it is 10 utterances away from the anaphor. As for coref-hoi-utt, the resolver successfully extracts the anaphor but incorrectly selects "Its just two point five for that one" as the antecedent, which, like the antecedent chosen by coref-hoi, is farther away from the anaphor than the correct antecedent is. We speculate that coref-hoi and coref-hoi-utt fail to identify the correct antecedent because they do not explicitly model distance and therefore may not have an idea about how far a candidate antecedent is from the anaphor under consideration. The additional features that dd-utt has access to, including the features that encode sentence distance as well as those that capture contextual information, may have helped dd-utt choose the correct antecedent, but additional analysis is needed to determine the reason.

A: *You want your rating to be a two?*
A: Is that what you're saying?
B: Yeah, I just got it the other way.
B: Uh in Yep, I just got
A: Okay.
A: So, I'll work out the average for that again at the end.
A: It's very slightly altered. Okay, and we're just waiting for your rating.
B: two point five
C: *Its just two point five for that one.*
A: Two point five, okay.
D: Yeah.
A: **Losing one decimal place, *that* is okay.**

Figure 3: Example in which the correct antecedent is identified by dd-utt but not by the baseline resolvers.

7 Conclusion

We presented an end-to-end discourse deixis resolution model that augments Lee et al.'s (2018) span-based entity coreference model with 10 extensions. The resulting model achieved state-of-the-art results on the CODI-CRAC 2021 datasets. We employed a variant of this model in our recent participation in the discourse deixis track of the CODI-CRAC 2022 shared task (Yu et al., 2022a) and achieved the best results (see Li et al. (2022) for details). To facilitate replicability, we make our source code publicly available.[4]

[4] See our website at https://github.com/samlee946/EMNLP22-dd-utt for our source code.

Limitations

Below we discuss several limitations of our work.

Generalization to corpora with clausal antecedents. As mentioned in the introduction, the general discourse deixis resolution task involves resolving a deictic anaphor to a clausal antecedent. The fact that our resolver can only resolve anaphors to utterances raises the question of whether it can be applied to resolve deictic anaphors in texts where antecedents can be clauses. To apply our resolver to such datasets, all we need to do is to expand the set of candidate antecedents of an anaphor to include those clauses that precede it. While corpora annotated with clausal antecedents exist (e.g., TRAINS-91 and TRAINS-93), we note that the decision made by the CODI-CRAC 2021 shared task organizers to use utterances as the unit of annotation has to do with annotation quality, as the inter-annotator agreement on the selection of clausal antecedents tends to be low (Poesio and Artstein, 2008),

Discourse deixis resolution in dialogue vs. narrative text. Whether our model will generalize well to non-dialogue datasets (e.g., narrative text) is unclear. Given the differences between dialogue and non-dialogue datasets (e.g., the percentage of pronominal anaphors), we speculate that the performance of our resolver will take a hit when applied to resolving deictic anaphors in narrative text.

Size of training data. We believe that the performance of our resolver is currently limited in part by the small amount of data on which it was trained. The annotated corpora available for training a discourse deictic resolver is much smaller than those available for training an entity coreference resolver (e.g., OntoNotes (Hovy et al., 2006)).

Data biases. Generally, our work should not cause any significant risks. However, language varieties not present in training data can potentially amplify existing inequalities and contribute to misunderstandings.

Acknowledgments

We thank the three anonymous reviewers for their insightful comments on an earlier draft of the paper. 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*, pages 32–42, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the LREC Workshop on Linguistics Coreference*, pages 563–566.

Donna K. Byron. 2002. Resolving pronominal reference to abstract entities. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 80–87, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Miriam Eckert and Michael Strube. 2000. Dialogue acts, synchronizing units, and anaphora resolution. *Journal of Semantics*, 17:51–89.

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

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

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

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

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

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

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

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.

Shengjie Li, Hideo Kobayashi, and Vincent Ng. 2021. The CODI-CRAC 2021 shared task on anaphora, bridging, and discourse deixis resolution in dialogue: A cross-team analysis. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 71–95, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shengjie Li, Hideo Kobayashi, and Vincent Ng. 2022. Neural anaphora resolution in dialogue revisited. In *Proceedings of the CODI-CRAC 2022 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 32–47, Gyeongju, Republic of Korea. 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.

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

Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, Sebastien Bourban, Mike Flynn, Maël Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Wilfried Post, Dennis Reidsma, and Pierre D. Wellner. 2005. The AMI meeting corpus. The AMI Project Consortium, `www.amiproject.org`.

Christoph Müller. 2008. *Fully Automatic Resolution of "it", "this", and "that" in Unrestricted Multi-Party Dialog*. Ph.D. thesis, Universität Tübingen, Tübingen, Germany.

Costanza Navarretta. 2000. Abstract anaphora resolution in Danish. In *1st SIGdial Workshop on Discourse and Dialogue*, pages 56–65, Hong Kong, China. Association for Computational Linguistics.

Eric W. Noreen. 1989. *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. John Wiley & Sons Inc.

Massimo Poesio and Ron Artstein. 2008. Anaphoric annotation in the ARRAU corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Michael Strube and Christoph Müller. 2003. A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 168–175, Sapporo, Japan. Association for Computational Linguistics.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 673–683, Hong Kong, China. 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*.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

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

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.

Juntao Yu, Sopan Khosla, Ramesh Manuvinakurike, Lori Levin, Vincent Ng, Massimo Poesio, Michael Strube, and Carolyn Rosé. 2022a. The CODI-CRAC 2022 shared task on anaphora, bridging, and discourse deixis in dialogue. In *Proceedings of the CODI-CRAC 2022 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 1–14, Gyeongju, Republic of Korea. Association for Computational Linguistics.

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

A Detailed Experimental Results

We report the resolution results of the four resolvers (UTD_NLP, coref-hoi, coref-hoi-utt, and dd-utt) on the CODI-CRAC 2021 shared task test sets in terms of MUC, B^3, and CEAF$_e$ scores in Table 8 and their mention extraction results in terms of recall (R), precision (P), and F-score (F) in Table 9.

Consider first the resolution results in Table 8. As can be seen, not only does dd-utt achieve the best CoNLL scores on all four datasets, but it does so via achieving the best MUC, B^3, and CEAF$_e$ F-scores. In terms of MUC F-score, the performance difference between dd-utt and the second best resolver on each dataset is substantial (2.2%–14.9% points). These results suggest that better link identification, which is what the MUC F-score reveals, is the primary reason for the superior performance of dd-utt. Moreover, Persuasion appears to be the easiest of the four datasets, as this is the dataset on which three of the four resolvers achieved the highest CoNLL scores. Note that Persuasion is also the dataset on which the differences in CoNLL score between dd-utt and the other resolvers are the smallest. These results seem to suggest that the performance gap between dd-utt and the other resolvers tends to widen as the difficulty of a dataset increases.

Next, consider the *anaphor* extraction results in Table 9. In terms of F-score, dd-utt lags behind UTD_NLP on two datasets, AMI and Switchboard. Nevertheless, the anaphor extraction *precision* achieved by dd-utt is often one of the highest in each dataset.

	MUC			B^3			CEAF$_e$			
	P	R	F	P	R	F	P	R	F	CoNLL
LIGHT										
UTD_NLP	44.6	31.3	36.8	56.2	37.0	44.6	55.3	40.5	46.7	42.7
coref-hoi	37.2	36.3	36.7	48.9	42.0	45.2	58.2	38.5	46.3	42.7
coref-hoi-utt	36.5	38.8	37.6	46.7	42.3	44.4	55.3	38.0	45.0	42.3
dd-utt	52.4	41.3	46.2	62.0	41.6	49.8	69.0	37.6	48.7	48.2
AMI										
UTD_NLP	45.5	21.2	28.9	52.4	29.5	37.8	44.9	35.1	39.4	35.4
coref-hoi	21.7	30.5	25.4	28.7	36.3	32.1	39.0	31.0	34.6	30.7
coref-hoi-utt	25.5	33.1	28.8	34.6	39.0	36.7	43.4	36.1	39.4	35.0
dd-utt	41.2	39.8	40.5	48.9	42.8	45.6	54.4	37.5	44.4	43.5
Persuasion										
UTD_NLP	45.5	20.3	28.1	65.0	30.2	41.2	61.0	41.8	49.6	39.6
coref-hoi	48.6	42.3	45.2	57.5	45.9	51.1	66.2	44.0	52.9	49.7
coref-hoi-utt	50.0	49.6	49.8	56.8	51.7	54.1	64.4	49.4	55.9	53.3
dd-utt	56.7	48.0	52.0	63.8	49.9	56.0	72.1	46.9	56.8	54.9
Switchboard										
UTD_NLP	35.2	21.3	26.5	52.3	30.4	38.5	50.5	34.9	41.3	35.4
coref-hoi	31.5	30.4	31.0	40.9	34.0	37.1	51.4	30.2	38.0	35.4
coref-hoi-utt	30.6	29.3	29.9	39.5	32.7	35.8	49.5	29.2	36.7	34.1
dd-utt	46.3	43.4	44.8	54.9	44.5	49.2	63.4	38.3	47.7	47.2

Table 8: Resolution results on the test sets.

		LIGHT			AMI			Persuasion			Switchboard		
		P	R	F	P	R	F	P	R	F	P	R	F
Overall	UTD_NLP	65.2	46.9	54.6	60.2	39.1	47.4	72.3	41.6	52.8	64.4	42.2	51.0
	coref-hoi	62.9	49.5	55.4	40.5	42.7	41.5	68.6	52.0	59.2	55.3	41.2	47.2
	coref-hoi-utt	59.3	50.0	54.2	43.9	45.2	44.5	66.2	57.6	61.6	53.3	39.6	45.5
	dd-utt	72.6	46.9	57.0	57.8	46.6	51.6	73.9	54.7	62.8	66.9	49.6	57.0
Anaphor	UTD_NLP	71.4	68.8	70.1	58.0	64.4	61.0	76.7	64.2	69.9	65.7	70.7	68.1
	coref-hoi	71.8	70.0	70.9	42.2	59.3	49.3	72.9	63.4	67.8	63.0	60.8	61.9
	coref-hoi-utt	68.2	72.5	70.3	46.4	60.2	52.4	71.3	70.7	71.0	61.9	59.3	60.6
	dd-utt	81.0	63.8	71.3	57.9	55.9	56.9	77.9	65.9	71.4	67.5	63.1	65.2
Antecedent	UTD_NLP	50.8	27.7	35.8	66.0	20.5	31.3	59.6	21.2	31.3	60.8	21.5	31.7
	coref-hoi	52.7	34.8	41.9	38.3	30.4	33.9	63.9	42.5	51.0	46.3	27.2	34.3
	coref-hoi-utt	49.4	33.9	40.2	41.0	34.2	37.3	60.7	46.6	52.7	43.3	25.5	32.1
	dd-utt	63.9	34.8	45.1	57.7	39.8	47.1	69.5	45.2	54.8	66.2	40.0	49.8

Table 9: Mention extraction results on the test sets.