
End-to-End Neural Discourse Deixis Resolution in Dialogue

Abstract

We adapt a 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.

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. 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. 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, which are arguably harder to encode than noun phrase semantics; and (2) string matching plays little role in DD resolution, unlike in entity coreference.

We focus on end-to-end DD resolution in dialogue. 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, the percentage of pronominal deictic anaphors rises to 93

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, a re-implementation of a span-based entity coreference model has been applied to resolve the deictic anaphors in the DD track after augmenting it with a type prediction model. Not only did they achieve the highest score on each dataset, but they beat the second-best system, 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 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. 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.

Table 1: Statistics on the datasets.

	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
LIGHT 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
AMI 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
Pers. 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
Swbd. test	22	3652	1996	166.0	9.6	90.7	12.0	14.7	2.0

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. 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. 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, a Siamese neural network is used, 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, a Transformer-based approach to DD resolution has recently been proposed, which is an end-to-end coreference system based on SpanBERT. Their model jointly learns mention extraction and DD resolution and has achieved state-of-the-art results.

3 Corpora

We use the DD-annotated corpora provided as part of the shared task. For training, we use the official training corpus from the shared task, ARRAU, 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, LIGHT, Persuasion, and Switchboard. Statistics on these corpora are provided in Table 1.

4 Baseline Systems

We employ three baseline systems.

The **first baseline**, coref-hoi, is a re-implementation of a 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 z , the model learns a distribution $P(y)$ over its antecedents $y \in \mathcal{Y}(z)$, where $\mathcal{Y}(z)$ includes a dummy antecedent ϵ and every preceding span:

$$P(y) = \frac{e^{s(z,y)}}{\sum_{y' \in \mathcal{Y}(z)} e^{s(z,y')}} \quad (1)$$

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 to

the same entity ($s_c(z, \epsilon) = s_a(z, \epsilon) = 0$ for dummy antecedents):

$$s(z, y) = s_m(x) + s_m(y) + s_c(z, y) + s_a(z, y) \quad (2)$$

$$s_m(\cdot) = \text{FFNN}_m(g_z) \quad (3)$$

$$s_c(z, y) = g_x^T W_c g_y \quad (4)$$

$$s_a(z, y) = \text{FFNN}_a([g_x, g_y, g_x \odot g_y, \phi(x, y)]) \quad (5)$$

where g_x and g_y are the vector representations of x and y , W_c is a learned weight matrix for bilinear scoring, $\text{FFNN}(\cdot)$ 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, is the top-performing system in the DD track of the shared task. 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.

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 datasets are all utterances.

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 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 $\gamma_1 \text{Dist}(x, y)$ from $s(x, y)$ (see Equation (1)), where $\text{Dist}(x, y)$ 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 $\gamma_2 \frac{1}{\text{Length}(y)}$ from $s(x, y)$, where $\text{Length}(y)$ 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 o_{ti} 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 o_{ti} is bigger than its second value:

$$o_{ti} = \text{FFNN}(g_i) \quad (6)$$

$$t_i = \arg \max_{x \in \{A, NA\}} o_{ti}(x) \quad (7)$$

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 o_{ti} , 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:

$$p_1(i) = \begin{cases} 0 & \text{if } \arg \max_{y \in \mathcal{Y}} s(i, y) = \epsilon \text{ and } t_i = NA \\ o_{ti}(A) - o_{ti}(NA) & \text{otherwise} \end{cases} \quad (8)$$

Intuitively, p_1 estimates the minimum amount to be adjusted so that span i 's type is not ANAPHOR. We incorporate p_i into the model as a penalty term in s (Equation (1)). Specifically, we redefine $s(i, \epsilon)$ as shown below:

$$s(i, \epsilon) = s(i, \epsilon) - [\gamma_3 p_1(i)] \quad (9)$$

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:

$$p_2(i) = \begin{cases} o_{ti}(NA) - o_{ti}(A) & \text{if } \arg \max_{y \in \mathcal{Y}} s(i, y) \neq \epsilon \text{ and } t_i = NA \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

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

$$s(i, j) = s(i, j) - [\gamma_4 p_2(i)] \quad (11)$$

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 $s_a(x, y)$:

$$s_c(x, y) = g_x^T W_c g_y + g_s^T W_a g_y \quad (12)$$

$$s_a(x, y) = \text{FFNN}_a([g_x, g_y, g_x \odot g_y, g_s, \phi(x, y)]) \quad (13)$$

Table 2: Lists of filtered words.

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, oops
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

where W_c and W_a 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.

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. 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, B³, and CEAF_e. In addition, we report the results of deictic anaphor recognition. We express recognition results in terms of Precision (P), Recall (R) and F-score, con-

Table 3: Resolution and recognition results on the four 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 4: Parameter values enabling dd-utt to achieve the best CoNLL score on each development set.

	LIGHT	AMI	Pers.	Swbd.
Type loss coef. 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

sidering 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 SpanBERTLarge as the encoder and reuse the hyperparameters 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; 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. For dd-utt, we use SpanBERTLarge 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., $\lambda, \gamma_1, \gamma_2, \gamma_3, \gamma_4$) 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}.

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 AMItest, we train models on ARRAU, LIGHTdev, Persuasiondev and Switchboarddev and use AMIdev 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 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. 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 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.

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, B3, and CEAF_e scores are reported in Table. Their mention extraction results in terms of recall (R), precision (P), and F-score (F) are provided in Table.

dd-utt achieves the best CoNLL scores on all four datasets, via achieving the best MUC, B3, 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.

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

7 Further Analysis

An example is analyzed. 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. 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

Table 5: Resolution results on the test sets.

	MUC			B3			CEAF _e			CoNLL
	P	R	F	P	R	F	P	R	F	
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	37.6	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

additional features that dd-utt has access to, including those that encode sentence distance as well as those that capture contextual information, may have helped dd-utt choose the correct antecedent.

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.

8 Error Analysis

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 "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

Table 6: Mention extraction results on the test sets.

	LIGHT			AMI			Persuasion		
	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
coref-hoi	62.9	49.5	55.4	40.5	42.7	41.5	68.6	52.0	59.2
coref-hoi-utt	59.3	50.0	54.2	43.9	45.2	44.5	66.2	57.6	61.6
dd-utt	72.6	46.9	57.0	57.8	46.6	51.6	73.9	54.7	62.8
Anaphor									
UTD_NLP	71.4	68.8	70.1	58.0	64.4	61.0	76.7	64.2	69.9
coref-hoi	71.8	70.0	70.9	42.2	59.3	49.3	72.9	63.4	67.8
coref-hoi-utt	68.2	72.5	70.3	46.4	60.2	52.4	71.3	70.7	71.0
dd-utt	81.0	63.8	71.3	57.9	55.9	56.9	77.9	65.9	71.4
Antecedent									
UTD_NLP	50.8	27.7	35.8	66.0	20.5	31.3	59.6	21.2	31.3
coref-hoi	52.7	34.8	41.9	38.3	30.4	33.9	63.9	42.5	51.0
coref-hoi-utt	49.4	33.9	40.2	41.0	34.2	37.3	60.7	46.6	52.7
dd-utt	63.9	34.8	45.1	57.7	39.8	47.1	69.5	45.2	54.8

Switchboard			
	P	R	F
Overall			
UTD_NLP	64.4	42.2	51.0
coref-hoi	55.3	41.2	47.2
coref-hoi-utt	53.3	39.6	45.5
dd-utt	66.9	49.6	57.0
Anaphor			
UTD_NLP	65.7	70.7	68.1
coref-hoi	63.0	60.8	61.9
coref-hoi-utt	61.9	59.3	60.6
dd-utt	67.5	63.1	65.2
Antecedent			
UTD_NLP	60.8	21.5	31.7
coref-hoi	46.3	27.2	34.3
coref-hoi-utt	43.3	25.5	32.1
dd-utt	66.2	40.0	49.8

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.

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.

9 Conclusion

An end-to-end discourse deixis resolution model that augments Lee et al.'s (2018) span-based entity coreference model with 10 extensions is presented. The resulting model achieved state-of-the-art results on the CODI-CRAC 2021 datasets.