Neural Discourse Deixis Resolution in Dialogue

Anonymous ACL submission

Abstract

We adapt Lee at el.'s (2018) span-based entity coreference model to the task of discourse deixis resolution. The resulting model 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 DD to its antecedent. A *deixis* is a term coined by Webber (1991) to denote a reference to a discourse entity such as a proposition, description, event, or speech act. As an example, consider the partial dialogue in Figure 1. In this example, the deixis "it" refers to the utterance by B in which s/he said s/he would donate $10. Unlike entity coreference, where lexical overlap is a strong indicator of coreference, commonly used deictic expressions (e.g., "that", "this", "it") are semantically empty and therefore cannot be resolved by simple string-matching facilities.

A natural question, then, is: how successful would a state-of-the-art entity coreference model be when applied to DD resolution? Recently, Kobayashi et al. (2021) applied Xu and Choi's (2020) re-implementation of Lee et al.'s span-based entity coreference model after augmenting it with a *type prediction* model (see Section 4.2) to resolve deictic expressions in the discourse deixis track of the CODI-CRAC 2021 shared task. Not only did they achieve the highest score on each dataset, 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 goal in this paper is to investigate whether *task-specific* observations can be exploited to extend a span-based model to further improve its performance for DD resolution. Empirical results on the CODI-CRAC 2021 shared task datasets show that our extensions, though simple, are surprisingly effective in improving model performance.

A: Would you donate to Save the Children?
B: {Yes, I will do $10 to both.}$_{\text{antecedent}}$
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}$_{\text{anaphor}}$.

Figure 1: Example

2 Related Work

Broadly, existing approaches to DD resolution can be divided into two categories, rule-based systems (e.g., Eckert and Strube (2000), Byron (2002) Navarretta (2000)) and machine learning-based systems (e.g., Strube and Müller (2003), Müller (2008)). Recently, deep learning is applied to the task, using Siamese neural networks (Marasović et al., 2017; Anikina et al., 2021) or span-based models (Kobayashi et al., 2021) to rank the candidate antecedents of a DD. See Appendix A for a detailed discussion of related work.

3 Corpora

We used 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, RST) and two non-dialogue sub-corpora (GNOME, PEAR). For validation and evaluation, we use the official development sets and test sets from the shared task. The corpus is composed of four well-known conversational datasets: the AMI corpus (McCowan et al., 2005), the LIGHT corpus (Urbanek et al., 2019), the Persuasion corpus (Wang et al., 2019), and Switchboard (Godfrey et al., 1992). Basic statistics about these corpora are provided in Table 1. Additional statistics can be found in Appendix B.

		Total #docs	Total #sents	Total #tokens	Total #ana	Total #ante
ARRAU	train	552	22406	348072	1624	2677
LIGHT	dev	20	908	11495	62	83
	test	21	923	11824	80	96
AMI	dev	7	4139	33741	230	294
	test	3	1967	18260	118	142
Pers.	dev	21	812	9185	94	94
	test	28	1139	12629	123	134
Swbd.	dev	11	1342	14992	127	175
	test	22	3652	35027	263	323

Table 1: Statistics on the datasets.

4 Two 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 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 pair-wise 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 ($s_a(x, \epsilon) = 0$ for dummy antecedents):

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

where g_x and g_y are the vector representations of x and y, W_c is a learned weight matrix for bilinear scoring, $\texttt{FFNN}(\cdot)$ 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 CODI-CRAC 2021 shared task (Kobayashi et al., 2021), which extends `coref-hoi` with a set of modifications. Two of the most important modifications are: (1) the addition of a sentence distance feature into $\phi(\cdot)$, 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 ANAPHORs can be resolved, and they can only be resolved to spans predicted as ANTECEDENTs. Details of how the type prediction model is trained can be found in Kobayashi et al. (2021).

5 Approach

In this section, we describe the extensions we make to the `UTD_NLP` model.

5.1 Candidate Anaphor Extraction

Our first extension, candidate anaphor extraction, is motivated by the observation that most deictic expressions are demonstrative pronouns (e.g., "that", "this") and "it". In our development sets, these three pronouns alone account for 84–88% of the anaphors. Consequently, we modify `UTD_NLP` as follows: instead of allowing each span of length n or less to be a candidate anaphor, we only allow a span in which the underlying word/phrase has appeared at least once in the training set to be a candidate anaphor. While this seems like a very simple extension, doing so substantially reduces the memory requirements of the model and enables the model to fit into memory even after it is augmented with all of the extensions that we will see in the rest of this section.

5.2 Anaphor Prediction

Now that we have the candidate anaphors, our second extension involves predicting which of these candidate anaphors are indeed deictic expressions. To do so, we retrain the type prediction model in `UTD_NLP` to predict the type of each candidate anaphor span. The type for a span i is ANAPHOR if i corresponds to a deictic expression, or NULL if it does not. Only spans that are predicted to be ANAPHOR will be resolved by the model.

5.3 Candidate Antecedent Extraction

A closer examination of our development sets reveals that only utterances can serve as the antecedents of deictic expressions. Thus, rather than enumerating all spans of up to a certain length to obtain the candidate antecedents, we only allow ut-

terances to be candidate antecedents. More specifically, we extract candidate antecedents as follows. For each span i that is predicted to be ANAPHOR by the type prediction model, we select the 10 utterances closest to i (including the utterance in which i appears) to be its candidate antecedents. The reasons are that (1) deictic expressions are anaphoric expressions, and hence recency plays an important role in antecedent selection, and (2) in our development sets, the 10 closest utterances already cover 96–99% of the antecedent-anaphor pairs.

5.4 Dummy Antecedent Elimination

Our next extension involves eliminating dummy candidate antecedents. Recall that in `coref-hoi`, the set of candidate antecedents for every span includes the dummy antecedent, which will be selected as the antecedent of a span i if (1) i is not an entity mention or (2) i is an entity mention but it is not anaphoric.

For our model, the situation is different. Since only those spans predicted as ANAPHOR by the anaphor prediction model described in Section 5.2 will be passed to the antecedent selection model, the antecedent selection model only sees spans that have been classified as anaphoric. Since these spans are anaphoric, they should presumably not be resolved to the dummy antecedent. For this reason, we eliminate the dummy antecedent from the set of candidate antecedents of every span when training and testing the antecedent selection model.

5.5 Features

Our next extension involves a large-scale expansion of features, hypothesizing that hand-engineered features could be profitably used by a span-based model. Specifically, we incorporate three types of features: (1) anaphor-based features, which encode the context of an anaphor, (2) antecedent-based features, which encode some statistics computed based on a candidate antecedent, and (3) pairwise features, which encode the relationship between an anaphor and a candidate antecedent. The list of features is shown in Table 2.

We add these features to both the bilinear score $s_c(x, y)$ and the concatenation-based $s_a(x, y)$:

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

where W_c and W_s are learned weight matrix, g_s is the embeddings of the sentence s that anaphor x is in, $\phi(x, y)$ encodes the speaker information as well as different types of distance between two spans.

Type	Features
Anaphor	Embedding of the sentence the anaphor is in
Antecedent	# of words; # of nouns; # of verbs; # of adjectives; # of content word overlaps between antecedent and the preceding words of the anaphor; whether an antecedent is the longest among all candidate antecedents; whether an antecedent has the most content word overlap among all candidate antecedents
Pairwise	Sentence distance between a candidate antecedent and an anaphor, ignoring sentences that contain only interjections, filling words, reporting verbs, and punctuation[1]

Table 2: Additional features used in our model.

5.6 Inference-Time-Only Distance-Based Candidate Antecedent Filtering

Given that we have fewer training instances for those antecedent-anaphor pairs that have larger sentence distances and it is generally harder to learn long-distance dependencies, correctly resolving an anaphor whose antecedent is far away from it is by no means easy. Although we used only the 10 closest utterances during training, we propose to further lower this number during inference. Specifically, for each candidate anaphor, the model selects an antecedent from one of the n closest utterances where $1 \leq n < 10$, where n is a tunable parameter.

6 Evaluation

6.1 Experimental Setup

Evaluation metrics. In the shared task, DD resolution is a generalized case of event coreference. Thus, resolution performance is evaluated using the well-known CoNLL score, which is the unweighted average of the F-scores from three metrics, MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), and $CEAF_e$ (Luo, 2005).

Model training and parameter tuning. For `coref-hoi`, we use $\text{SpanBERT}_{\text{Large}}$ as the encoder and reuse the hyperparameters from Xu and Choi (2020) with only one exception: we increase the maximum span width from 30 to 45, which covers more than 98% of the antecedent spans.

[1]Note that the sentence distance feature introduced by Kobayashi et al. (2021) is the *raw* sentence distance, whereas the one we introduce here is a refined sentence distance feature that aims to more accurately capture proximity by ignoring *meaningless* sentences (e.g., those that contain interjections). The complete list of filling words and reporting verbs that we filter can be found in Appendix C

	LIGHT	AMI	Pers.	Swbd.	Avg.
UTD_NLP^P	42.70	35.35	39.64	35.43	38.28
coref-hoi^P	39.40	24.95	33.00	32.03	32.34
Ours^P	**49.45**	**42.33**	**56.13**	**47.63**	**48.89**
UTD_NLP^G	43.44	36.91	52.09	40.44	43.22
coref-hoi^G	44.92	29.62	46.70	28.85	37.52
Ours^G	**45.97**	**39.62**	**53.76**	**51.11**	**47.62**

Table 3: Results on each of the four test sets. We report the CoNLL F score on each test set, as well as an average CoNLL F score over all test sets. Detailed MUC, B^3, $CEAF_e$ scores can be found in Appendix E

For `UTD_NLP`, we will simply report their official results on the shared task test sets.

For our model, we use $\text{SpanBERT}_{\text{Large}}$ as the encoder. We tune the type loss coefficient λ and distance-based candidate antecedent filtering parameter n using grid search to maximize the CoNLL score on dev data. Since we do not rely on span enumerate to generate candidate spans, the maximum span width can be set to any arbitrary numbers that are large enough to cover all our candidate antecedents and anaphors. In our case, we use 300 as our maximum span width. We reuse other hyperparameters from Xu and Choi (2020). The search range of each hyperparameter and the final hyperparameters are listed in Appendix D.

Both `coref-hoi` and our model are trained for 30 epochs with a dropout rate of 0.3 and early stopping.

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 development. For example, when evaluating on AMI_{test}, we train models on ARRAU, $\text{LIGHT}_{\text{dev}}$, $\text{Persuasion}_{\text{dev}}$ and $\text{Switchboard}_{\text{dev}}$ and use AMI_{dev} for development.

6.2 Results

We follow the shared task to obtain results in two different settings: (1) the Predicted setting, where models need to extract anaphor mentions from the documents, and (2) the Gold setting, where models need to extract anaphor mentions from a given list of gold mentions. Results of the two baselines and our model on the four test sets are shown in Table 3, where the ^{P} and ^{G} models are respectively the models evaluated in the Predicted and Gold settings. As noted before, the results of `UTD_NLP` (the top-performing system in the shared task) are their official results in the shared task. 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.

	LIGHT	AMI	Pers.	Swbd.	Avg.
Ours^P−Features	44.58	44.12	50.98	49.70	47.34
Ours^P+Dummy	47.07	40.56	54.02	44.96	46.65
Ours^P,Dist10	49.45	42.30	55.08	47.21	48.51

Table 4: Ablation results for the Predicted setting.

Our models `Ours`^{P} and `Ours`^{G} outperform all other models on every test set in both settings and the gold setting, despite the unfair advantage that `UTD_NLP` has in terms of parameter tuning. Overall, `Ours`^{P} improves the previous state-of-the-art results in the predicted setting by 10.6 CoNLL score, and `Ours`^{G} outperforms its previous state-of-the-art counterpart by 4.4 CoNLL score.

6.3 Ablation Results

To gain insights into `Ours`^{P}, we conduct three ablation experiments[2]:

(1) Ablating features (`Ours`^{P}`−Feats`): We remove all but the three features used in the `UTD_NLP` baseline, so that we can evaluate the benefits of having the features in Table 2.

(2) Including dummy antecedents (`Ours`^{P}`+Dummy`): We add back the dummy antecedent as a candidate antecedent for each candidate anaphor and re-train the model, so that we can evaluate how much we gained by removing dummy antecedents.

(3) Removing distance-based filtering (`Ours`^{P}`,Dist10`): We set the sentence distance filtering parameter n to 10, which is equivalent to not performing sentence distance-based filtering.

Results are shown in Table 4. Ablating the features causes a 1.55 drop in CoNLL score adding back the dummy antecedent causes a 2.34 drop, and not doing sentence distance filtering causes a 0.38 drop. Note that even these ablated models outperform their current state-of-the-art counterpart `UTD_NLP`^{P} by 8.37–10.23 in CoNLL score.

7 Conclusion

We presented an end-to-end discourse deixis resolution model that achieves state-of-the-art results on the CODI-CRAC 2021 datasets.

[2]Ablation results for the Gold setting are in Appendix F.

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 *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566.

Donna K. Byron. 2002. Resolving pronominal reference to abstract entities. In *ACL*.

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

J.J. Godfrey, E.C. Holliman, and J. McDaniel. 1992. Switchboard: telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520 vol.1.

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.

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 *EMNLP*.

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.

M. Müller. 2008. Fully automatic resolution of 'it', 'this', and 'that' in unrestricted multi-party dialog.

Costanza Navarretta. 2000. Abstract anaphora resolution in danish. In *SIGDIAL Workshop*.

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 M. Müller. 2003. A machine learning approach to pronoun resolution in spoken dialogue. In *ACL*.

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.

A Further Discussion of Related Work

Rule-based. Early works that tackle the resolution of discourse deixis mentions are rule-based systems (Eckert and Strube, 2000; Byron, 2002; Navarretta, 2000). They use predefined rules to extract anaphor mentions, and select antecedent for each extracted anaphor based on the dialogue act types of each candidate antecedent.

Machine learning-based. Early machine learning works (Strube and Müller, 2003; Müller, 2008) use hand-crafted feature vectors to represent mentions. A classifier is then trained to identify whether a pair of mentions is a valid antecedent-anaphor pair. Recent machine learning approaches (Marasović et al., 2017; Anikina et al., 2021) use a Siamese neural network to rank candidate antecedents of an anaphor. They use as their model input the sentence embeddings obtained from a bi-LSTM model or a BERT model. A score is then generated for each pair of input sentences by the Siamese neural network, and is later used to rank the candidate antecedents.

Transformers-based. Transformers-based entity coreference resolution approaches have emerged several years ago (Kantor and Globerson, 2019; Joshi et al., 2019, 2020). However, transformers-based discourse deixis resolution approach is only recently proposed by Kobayashi et al. (2021), which is an end-to-end coreference system based on SpanBERT. Their model jointly learns mention extraction and discourse deixis resolution, and achieves state-of-the-art results in the CODI-CRAC 2021 shared task.

B Detailed Statistics about the Corpus

Additional statistics about the corpus are shown in Table 5. Some interesting facts are worth mentioning. Firstly, the AMI datasets have 4 speakers per document on average. Secondly, the average number of anaphors and antecedents per document differ a lot between AMI_{dev} and AMI_{test}. These two factors can partially explain why all models perform the worst on AMI. Thirdly, the average number of anaphors and antecedents are very close on Persuasion, which means split-antecedents are very rare on Persuasion. This can be one of the reasons that all models perform the best on Persuasion.

C Complete List of Filtered Words

As stated in Table 2, we ignore some sentences that contain only certain words. The complete lists of words we ignore are as follows:

• 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.

D Hyperparameters and Experiment Details

For $\texttt{coref-hoi}^P$, $\texttt{coref-hoi}^G$, $\texttt{Ours}^P$, and $\texttt{Ours}^G$, we use 1×10^{-5} as our BERT learning rate, and 3×10^{-4} as our task learning rate. For $\texttt{Ours}^P$, and $\texttt{Ours}^G$, type loss coefficient of {0.2, 0.5, 1, 200, 500, 800} were tested, and sentence filtering parameter n of {1, 2, 3, 4, 5, 6, 7} were tested. Best hyperparameters were found using grid search. The final type loss coefficients and sentence filtering n's are shown in Table 6. All experiments were run using a random seed of 11.

		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 5: Additional statistics about the corpus. Turn information and speaker information are not annotated on the ARRAU dataset.

		LIGHT	AMI	Pers.	Swbd.
Ours^P	Type loss coefficient λ	0.5	0.5	0.5	0.5
	Sentence filtering n	10	6	3	5
Ours^G	Type loss coefficient λ	0.5	0.5	0.5	0.5
	Sentence filtering n	10	6	3	5

Table 6: The final type loss coefficients and sentence filtering n's of our models.

E Detailed Experiment Results

The detailed MUC, B^3, and $CEAF_e$ scores of each model are shown in Table 7. We also report the mention extraction results of each model in Table 8.

F Additional Ablation Experiments

We reported the ablation results of our models only in the predicted setting in Section 6.3. For completeness, we report the ablation results of our models in the gold setting in Table 9.

G Distribution of Antecedents over Sentence Distance

Table 10 shows the distribution of sentence distances between antecedents and anaphors in the development sets. As we can see, few antecedents are separated from their anaphors by more than a sentence distance of 5.

H Limitations of Our Work

We believe that the performance of our models is limited by the small amount of training data available. The corpora that are annotated for training a DD resolver is much small than the corpora annotated for training an entity coreference resolver.

	MUC			B^3			$CEAF_e$			
	P	R	F	P	R	F	P	R	F	CoNLL
					LIGHT					
UTD_NLP^P	44.6	31.2	36.8	56.2	37.0	44.6	55.3	40.5	46.7	42.7
coref-hoi^P	40.0	30.0	34.3	53.5	34.3	41.8	63.9	31.4	42.1	39.4
Ours^P	54.7	43.8	48.6	63.2	42.8	51.0	69.1	37.6	48.7	49.5
UTD_NLP^G	49.0	30.0	37.2	56.3	39.1	46.2	51.7	42.9	46.9	43.4
coref-hoi^G	55.8	36.2	43.9	66.0	35.9	46.5	72.9	31.9	44.3	44.9
Ours^G	48.4	37.5	42.3	60.9	40.3	48.5	70.8	35.4	47.2	46.0
					AMI					
UTD_NLP^P	45.5	21.2	28.9	52.4	29.5	37.8	44.9	35.1	39.4	35.4
coref-hoi^P	27.1	16.1	20.2	39.1	21.3	27.5	48.1	18.8	27.1	24.9
Ours^P	34.7	42.4	38.2	42.6	46.2	44.3	50.0	40.1	44.5	42.3
UTD_NLP^G	44.6	21.2	28.7	49.7	34.6	40.8	39.6	43.0	41.2	36.9
coref-hoi^G	24.1	23.7	23.9	34.0	30.8	32.3	48.0	24.7	32.6	29.6
Ours^G	33.3	36.4	34.8	42.1	41.9	42.0	49.7	36.4	42.0	39.6
					Persuasion					
UTD_NLP^P	45.5	20.3	28.1	64.9	30.2	41.2	61.0	41.8	49.6	39.6
coref-hoi^P	57.4	22.0	31.8	65.4	22.2	33.1	73.6	22.2	34.1	33.0
Ours^P	52.8	52.8	52.8	59.3	55.3	57.2	65.9	52.3	58.3	56.1
UTD_NLP^G	53.3	45.5	49.1	54.9	55.7	55.3	46.0	59.3	51.8	52.1
coref-hoi^G	56.4	35.8	43.8	64.5	37.5	47.5	72.0	37.0	48.9	46.7
Ours^G	57.1	45.5	50.7	64.5	47.7	54.8	71.6	45.6	55.8	53.8
					Switchboard					
UTD_NLP^P	35.2	21.3	26.5	52.3	30.4	38.5	50.5	34.9	41.3	35.4
coref-hoi^P	39.0	24.3	30.0	48.9	26.5	34.4	62.9	21.2	31.7	32.0
Ours^P	43.2	46.0	44.6	51.7	47.5	49.5	60.5	40.9	48.8	47.6
UTD_NLP^G	39.4	31.2	34.8	41.6	48.5	44.8	33.7	55.0	41.8	40.4
coref-hoi^G	42.4	20.2	27.3	53.0	21.4	30.5	63.3	18.6	28.7	28.8
Ours^G	47.6	49.4	48.5	56.1	50.6	53.2	63.6	43.4	51.6	51.1

Table 7: Detailed results for each model.

		Light			AMI			Persuasion			Switchboard		
		P	R	F	P	R	F	P	R	F	P	R	F
Anaphor	UTD_NLP^P	-	73.8	-	-	64.4	-	-	65.9	-	-	71.1	-
	coref-hoi^P	81.7	61.3	70.0	58.6	34.7	43.6	83.0	31.7	45.9	75.6	47.1	58.1
	Ours^P	78.1	62.5	69.4	53.5	65.3	58.8	71.5	71.5	71.5	64.3	68.4	66.3
	UTD_NLP^G	65.0	65.0	65.0	57.9	61.9	59.8	73.6	77.2	75.4	64.8	74.9	69.5
	coref-hoi^G	86.5	56.2	68.2	57.8	56.8	57.3	83.3	52.8	64.7	76.8	36.5	49.5
	Ours^G	82.3	63.7	71.8	52.7	57.6	55.1	78.6	62.6	69.7	68.5	71.1	69.8
Antecedent	UTD_NLP^P	-	27.7	-	-	20.5	-	-	21.2	-	-	21.5	-
	coref-hoi^P	56.4	27.7	37.1	47.6	18.6	26.8	63.6	19.2	29.5	62.9	21.2	31.7
	Ours^P	68.9	37.5	48.6	51.2	41.0	45.5	63.8	50.7	56.5	62.2	42.1	50.2
	UTD_NLP^G	59.7	33.0	42.5	49.5	32.3	39.1	52.4	58.9	55.5	38.2	52.4	44.2
	coref-hoi^G	67.3	29.5	41.0	47.0	24.2	32.0	65.3	33.6	44.3	58.3	17.1	26.5
	Ours^G	69.6	34.8	46.4	54.2	39.8	45.9	67.7	43.2	52.7	66.5	45.4	54.0

Table 8: Mention extraction results of each model. The precision and recall for the UTD_NLP^P model were not released by Kobayashi et al. (2021).

	LIGHT	AMI	Pers.	Swbd.	Avg.
Ours^G−Features	42.96	41.96	54.79	52.71	48.11
Ours^G+Dummy	45.13	42.22	52.34	50.68	47.59
Ours^G,Dist10	44.93	41.93	48.88	50.32	46.51

Table 9: Ablation results for gold setting.

		0	1	2	3	4	5	6	7	8	9	>9
LIGHT	dev	10	22	9	3	1	1	0	0	0	2	2
AMI	dev	32	65	41	19	8	8	4	1	1	0	4
Pers.	dev	13	64	14	0	1	0	0	0	0	0	1
Swbd.	dev	2	31	33	10	6	3	0	2	0	0	1

Table 10: The distribution of sentence distances between antecedent-anaphor pairs. Sentence distance of 0 means the anaphor and the antecedent are in the same sentence.