Online Coreference Resolution for Dialogue Processing: Improving Mention-Linking on Real-Time Conversations

Liyan Xu **Jinho D. Choi**
Department of Computer Science
Emory University, Atlanta, USA
{liyan.xu,jinho.choi}@emory.edu

Abstract

This paper suggests a direction of coreference resolution for online decoding on actively generated input such as dialogue, where the model accepts an utterance and its past context, then finds mentions in the current utterance as well as their referents, upon each dialogue turn. A baseline and four incremental-updated models adapted from the mention-linking paradigm are proposed for this new setting, which address different aspects including the singletons, speaker-grounded encoding and cross-turn mention contextualization. Our approach is assessed on three datasets: *Friends*, *OntoNotes*, and *BOLT*. Results show that each aspect brings out steady improvement, and our best models outperform the baseline by over 10%, presenting an effective system for this setting. Further analysis highlights the task characteristics, such as the significance of addressing the mention recall.

1 Introduction

It has been made practical recently to apply coreference resolution to assist a broad scope of NLP tasks (Peng et al., 2017; Sahu et al., 2019; Gao et al., 2019), especially with the advent of neural end-to-end decoding and contextualized encoding (Lee et al., 2017, 2018; Joshi et al., 2019, 2020; Wu et al., 2020). However, it is quite limited to use existing coreference models in real-time dialogue processing systems, as most of them are not trained to handle an online decoding environment. In the dialogue domain, recent efforts have focused on ellipsis recovery and query rewriting (Quan et al., 2019; Tseng et al., 2021); in this work, we target to address a new perspective specifically for the online decoding, where the model sequentially accepts utterances in a dialogue and spits out valid mentions as well as their referent links for each latest utterance turn upon arrival, to be consumed by the downstream dialogue processing (Figure 1).

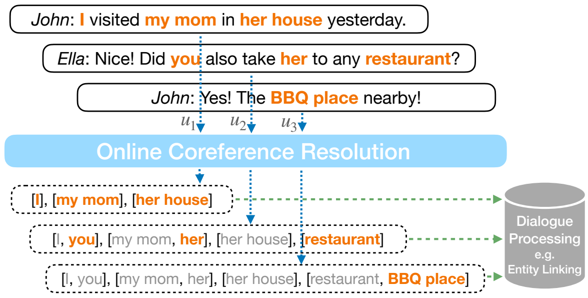

Figure 1: Illustration of the online setting. Predictions upon each turn are made immediately and ready for consumption by downstream applications. New mentions at each turn are marked by boldface in orange.

More formally, let u_i be the current (i'th) utterance in a dialogue $(u_1, .., u_i, ..)$; $\mathcal{M}_i$ be the mentions in u_i; $\mathcal{M}^{i-1}$ be the mentions from previously predicted clusters till u_{i-1}. The objective upon i'th turn is to: (1) identify $\mathcal{M}_i$ (2) identify conference links among $\mathcal{M}_i$, as well as from $\mathcal{M}_i$ to $\mathcal{M}^{i-1}$. We do not allow updates on $\mathcal{M}_i$ later, since that would be equivalent to general coreference resolution; in this work, we specifically target this underexplored online scenario under this setting, which requires accurate predictions upon each turn that could be directly consumed by downstream applications.

Several quasi-online coreference models have been proposed that maintain and update referents sequentially (Clark and Manning, 2015, 2016; Liu et al., 2019; Toshniwal et al., 2020; Xia et al., 2020). However, these models differ from our real online setting in two ways. First, only the latest utterance and its past sequence are visible in our setting, so that decisions need to be made without knowing the unseen future. Second, the decision of whether a span should be extracted or linked to others needs to be made immediately at each utterance turn, while quasi-online models can maintain an internal pool of candidates and make one final prediction after the entire document is processed.

For this task, we first introduce our baseline adapted from the classic mention-linking (ML) ap-

Proceedings of the 11th Joint Conference on Lexical and Computational Semantics, pages 341 - 347
July 14-15, 2022 ©2022 Association for Computational Linguistics

proach (Wiseman et al., 2015; Lee et al., 2017), and then propose four models where each one does an incremental update upon the previous model and addresses a specific perspective of this task, including the online inference, singletons, speaker-grounded encoding, and mention contextualization across utterance turns. For our approach, we do not use models that maintain explicit entities, because: (1) it has been shown that higher-order features from entity representation provide negative to marginal positive impact over ML counterparts despite their complexities (Xu and Choi, 2020; Xia et al., 2020; Toshniwal et al., 2020); (2) ML models are "stateless" so that they do not need to maintain decision states for previous mentions, which makes it more adaptable to applications in practice.

All models are evaluated on three datasets to test the generalizability of our approach, and the best model obtains over 10% improvement over the baseline on all datasets. Results and further analysis suggest that each aforementioned aspect can bring out steady improvement under the online setting, and highlight the singleton recovery to be the most critical component.

2 Approach

End-to-End Resolution Our model backbone is based on the end-to-end coreference resolution (Lee et al., 2018) with a Transformers encoder (Joshi et al., 2020). It scores every span for being a mention, and extracts top spans as mention candidates. Pairwise scoring is then performed among all candidates to determine the coreference links. Details of the model architecture can be referred by the paper from Lee et al. (2018), and we denote the original **c**oreference loss as $\mathcal{L}_c$.

Baseline (`BL`) We first present our baseline that takes the end-to-end model and trains in the exact same non-online way as prior work, but adapts the decoding to fit in our online inference setting.

Let u_i be the i'th utterance in the dialogue, and $|u_i|$ be its length (number of tokens). During online decoding upon u_i, this model takes an utterance sequence with past context as input, denoted by $\mathcal{U}_k^i = (u_k, .., u_i)$; $k \in [1, i)$ is dynamically determined by $\sum_{j=k}^{i} |u_j| \leq \Upsilon$ where Υ is the max number of tokens that the encoder accepts. Different from Lee et al. (2018), the mention candidates now consist of two parts: (1) the extracted top candidates solely from u_i, denoted as $\mathcal{X}_i$; (2) mentions from previously predicted clusters from $\mathcal{U}_k^{i-1}$, denoted as $\mathcal{M}_k^{i-1}$. Thereby the final candidate set $\mathcal{X}$ can be denoted as $\mathcal{X}_i \cup \mathcal{M}_k^{i-1}$. The same pairwise scoring as prior work is then performed on all candidates $\mathcal{X}$. Since we do not modify previous decisions in our setting, we keep coreference links among $\mathcal{X}_i$, or from $\mathcal{X}_i$ to $\mathcal{M}_k^{i-1}$, but not among $\mathcal{M}_k^{i-1}$. The predicted clusters after u_i will be updated in the same way by picking the referent antecedents according to coreference links.

Singleton Recovery (`SR`) `SR` is built upon `BL` to address the singleton problem. In `BL`, after processing each utterance sequence $\mathcal{U}_k^i$, the model filters out mention candidates from $\mathcal{X}_i$ that are not referent to any other candidates, according to the mention-linking paradigm. However, it results on losing non-anaphoric mentions that do not have referents in u_i, and yields a critical issue for online inference because mentions in u_i that are currently singletons but potentially will find referents in later utterances can get discarded too early.

To address this issue, we adopt a simple strategy similar to (Xu and Choi, 2021) that preserves any candidates whose mention scores are larger than a threshold of 0, denoted as $s_m > 0$, and creates a singleton cluster for each of which have not yet found any referent (intermediate singletons). However, as many annotation schemes do not require annotating singletons, e.g. CoNLL 2012, we may not have "true" gold labels covering every valid mentions, similar to the "misguidance of unlabeled entities" problem in named entity recognition (NER) (Li et al., 2021). Let Ψ_m^+ be the set of s_m of gold candidates according to the annotation, and Ψ_m^- be the set of s_m of other candidates that may also contain certain valid mentions (singletons). We mitigate the false negative issue of unlabeled mentions by applying dynamic negative sampling on Ψ_m^-, denoted as Φ_m^-, where $|\Psi_m^+| \approx |\Phi_m^-|$. Binary cross-entropy (`BCE`) loss is then used for this optimization to aid the threshold requirement:

$$\mathcal{L}_m = \texttt{BCE}(\Psi_m^+, \Phi_m^-) \quad (1)$$

$$\mathcal{L} = \alpha_c \cdot \mathcal{L}_c + \alpha_m \cdot \mathcal{L}_m \quad (2)$$

The final loss $\mathcal{L}$ is estimated by the weighted sum of $\mathcal{L}_m$ and $\mathcal{L}_c$ using the hyperparameters α_c and α_m.

Online Resolution (`OR`) `OR` is designed specifically for online inference on dialogues. Distinguished from `BL` that takes the whole document as input in training, `OR` takes $\mathcal{U}_k^i$ as input for both training and decoding, closing the gap. To capture subtle nuances from different speakers in the dialogue, we collect speaker names within each dialogue and assign a special token of position-based ID to each speaker (e.g. $\texttt{S}_1$, $\texttt{S}_2$) based on speaking orders, which is then prepended to its corresponding utterance (Wu et al., 2020). We also add `[SEP]` before u_i to signal the latest utterance. The following sequence is used as input for `OR`:

$$\{\texttt{S}_k\}^\frown u_k^\frown \cdots ^\frown \{\texttt{[SEP]}\}^\frown \{\texttt{S}_i\}^\frown u_i \quad (3)$$

During training upon the i'th turn, gold mentions in $\mathcal{U}_k^{i-1}$ are used as $\mathcal{M}_k^{i-1}$; the losses $\mathcal{L}_m$ and $\mathcal{L}_c$ are estimated only on candidates from u_i. Gradient accumulation is applied across multiple utterance turns, and we warm-start `OR` by initializing from the parameters of `SR`, followed by the online training described above. The decoding step for `OR` is kept the same as `BL` and `SR`.

Speaker-Grounding (`SG`) `SG` adds a speaker-grounding subtask upon `OR`, which is to facilitate the encoding of multi-speaker interaction which is an important aspect in dialogues. In `OR`, although each input token is conditioned on speaker tokens as in Eq (3), it is not obvious to the model that each token is from which speaker, which can be a barrier to learn the speaker interaction. To explicitly regularize the speaker encoding, we add a subtask to predict whether two candidates are from the same speaker based on their embeddings: the model gives a same-speaker score s_s such that pairs from the same speaker have $s_s > 0$ and others $s_s \leq 0$, forcing the semantic representation to fuse the speaker interaction. Let Ψ_s^+ be the set of s_s of pairs from the same speaker; Ψ_s^- be the set of s_s of other pairs. We optimize s_s by `BCE`, adding the loss in addition to $\mathcal{L}_c$ and $\mathcal{L}_m$:

$$s_s(x, y) = w_s \cdot [g_x \oplus g_y \oplus (g_x \circ g_y) \oplus (g_x - g_y)]$$

$$\mathcal{L}_s = \texttt{BCE}(\Psi_s^+, \Psi_s^-) \quad (4)$$

$$\mathcal{L} = \alpha_c \cdot \mathcal{L}_c + \alpha_m \cdot \mathcal{L}_m + \alpha_s \cdot \mathcal{L}_s \quad (5)$$

g_x/g_y denotes the representation of a candidate and w_s is the scoring parameter. $\oplus$ denotes concatenation and $\circ$ is the element-wise multiplication. We also apply negative sampling to keep $|\Psi_s^+| \approx |\Psi_s^-|$.

Span-Level Self-Attention (`SA`) `SA` is also added upon `OR` to achieve candidate contextualization. For each input $\mathcal{U}_k^i$, the representation of all candidates $\mathcal{X}$ is contextualized on the token-level because of Transformers' encoding. However, $\mathcal{M}_k^{i-1}$ is not used until the pairwise scoring. Therefore, $\mathcal{X}_i$ is not explicitly conditioned on the previously extracted mentions ($\mathcal{M}_k^{i-1}$) on the span-level. To capture the dependency among all mention candidates across utterances, we pass $\mathcal{X}$ to a scaled dot-product self-attention layer (Vaswani et al., 2017) before the pairwise scoring:

$$G' = \text{softmax}\Big(\frac{(GW_q)(GW_k)^T}{\sqrt{d}}\Big)(GW_v), \quad (6)$$

where $G \in \mathbb{R}^{|\mathcal{X}| \times d}$ is the embedding matrix of all candidates, d is the embedding size, W_q, W_k, W_v are the parameters. G' is the new candidate-aware embedding matrix, which provides enhanced candidate representation for the pairwise scoring.

3 Experiments

Datasets All models are experimented on the following three datasets. *Friends* contains transcripts from the TV show in which personal mentions are annotated for entity linking. Each scene is considered an independent dialogue where utterances and speaker IDs are provided. We adapt the data split suggested by Zhou and Choi (2018). *Onto-Conv* consists of documents in three genres selected from OntoNotes 5.0: broadcasting and telephone conversations, and web text including discussion forums. We adapt the data split provided by Pradhan et al. (2012) and treat each document as a dialogue and every sentence as an utterance. *BOLT* follows the same annotation guideline as OntoNotes although documents are from discussion forums, SNS chats, and telephone conversations (Li et al., 2016). Since this is the first work using *BOLT* for this task, we create a new data split for future replicability (see A.1). Out of these three datasets, only *Friends* provides annotation of singletons.

The numbers of documents in the training, development, and test set of *Friends*, *Onto-Conv*, *BOLT* are provided in Table 2, along with the averaged numbers of speakers, entity clusters and utterances per document of each dataset. More details regarding the datasets are provided in Appendix A.1.

Settings Our implementation are based on the PyTorch coreference models from Xu and Choi (2020), and SpanBERT$_{\text{BASE}}$ is adopted as the encoder. The implementation and trained models

	Friends				*Onto-Conv*				*BOLT*			
	MUC	B^3	$CEAF_{\phi_4}$	Avg F1	MUC	B^3	$CEAF_{\phi_4}$	Avg F1	MUC	B^3	$CEAF_{\phi_4}$	Avg F1
`BL`	81.9	62.2	54.5	66.2 (± 0.7)	70.5	54.8	43.9	56.4 (± 0.2)	73.3	61.2	51.1	61.9 (± 0.3)
`SR`	85.5	68.3	61.7	71.8 (± 0.5)	77.5	63.2	55.2	65.2 (± 0.6)	79.6	71.8	61.7	71.0 (± 0.4)
`OR`	85.8	71.9	65.7	74.5 (± 0.5)	78.0	63.6	55.6	65.7 (± 0.3)	79.5	72.0	63.2	71.5 (± 0.3)
`+SG`	85.7	73.6	67.0	75.3 (± 0.4)	78.1	**64.3**	56.5	66.3 (± 0.3)	**79.9**	72.3	63.4	71.8 (± 0.3)
`+SG+SA`	**86.4**	**73.7**	**68.2**	**76.1** (± 0.1)	**78.9**	**64.3**	**56.9**	**66.8** (± 0.1)	**79.9**	**72.7**	**64.1**	**72.3** (± 0.2)

Table 1: Results of all models in Section 2 on the evaluation sets of *Friends*, *Onto-Conv*, and *BOLT* datasets. MUC, B^3, and $CEAF_{\phi_4}$ show the F1 scores of the corresponding metrics, and their macro-average score (Avg F1) is used as the main evaluation metric. All scores presented here are the averaged scores over 3 repeated experiments; the standard deviations of Avg F1 scores are provided in the parentheses.

	TRN	DEV	TST	NS	NC	NU
`F`	987	122	192	3.7	4.6	18.7
`O`	566	100	95	2.4	16.2	49.5
`B`	943	117	117	2.9	9.2	18.1

Table 2: Statistics of the dataset *Friends* (`F`), *Onto-Conv* (`O`), *BOLT* (`B`). TRN, DEV, TST are the numbers of documents in the training, development, and test set of each dataset. NS, NC, NU are the averaged numbers of speakers, entity clusters, utterances per document of each dataset.

have been partially integrated with the open source project ELIT[1] (He et al., 2021).

During inference, all predicted clusters are collected and merged accordingly across utterances, and get evaluated by comparing them to the ground truth (all gold non-singleton clusters) at the end of each dialogue, in the same way as the CoNLL'12 shared task protocol. Detailed experimental settings are provided in Appendix A.2.

Results Table 1 describes the performance of all models on the test sets in the three datasets. These results are averaged across 3 repeated experiments; Avg-F1 is used as the main evaluation metric. Each proposed model gives steady improvement, and the best result is achieved by the `OR+SG+SA` model, surpassing the `BL` model on all datasets by significant margins of ≈10%. Among these models, singleton recovery contributes the most upon `BL`, demonstrating that albeit simple and intuitive, the training and inference of intermediate singletons is essential in online coreference resolution.

3.1 Analysis on Online Inference

To identify how model predictions are affected by online inference, all mentions in the predicted clusters are examined against the gold clusters. Table 3 shows the results of mention precision and recall from the four experimental settings.

	Friends		*Onto-Conv*		*BOLT*	
	P	R	P	R	P	R
`N:BL`	92.0	92.5	88.1	**83.6**	85.2	**82.8**
`O:BL`	**92.5**	85.3	**92.1**	60.6	**89.0**	64.8
`O:SR`	**92.5**	**93.2**	89.4	78.8	87.4	78.3
`O:SR-`	**92.5**	92.5	90.4	74.8	88.4	76.7

Table 3: The **P**recision and **R**ecall of all mentions in the predicted clusters on the test sets in the three datasets. `N` is **N**on-online inference as in CoNLL'12 shared task, `O` is **O**nline inference as in this work. `SR-` is the **S**ingleton **R**ecovery (`SR`) model without applying negative sampling on the mention loss in training.

Following observations are drawn by this analysis: **(1)** Comparing `N:BL` and `O:BL`, online inference indeed leads to a large drop on the mention recall as expected, without as much increase on precision, due to the omission of intermediate singletons. **(2)** Comparing `O:BL` and `O:SR`, singleton recovery (`SR`) significantly improves the mention recall (8% for *Friends* and 13+% for others) without sacrificing much precision. However, notice that the recall of `O:SR` for *Friends* is even higher than that of non-online inference (`N:BL`), but the recall for *Onto-Conv* and *BOLT* is still 4+% lower than that of `N:BL`. This is due to the fact that *Friends* does have singletons annotated while the other two do not. Thus, `O:SR` for *Friends* does not suffer from the "misguidance of unlabeled entities" problem. **(3)** Comparing `O:SR` and `O:SR-`, it illustrates the positive impact of applying negative sampling on mentions to alleviate the false-negative issue of unlabeled mentions, which improves recall while maintaining similar precision for online inference.

3.2 Analysis on Utterance Interaction

As we aim to build a robust online resolution model in the dialogue domain, understanding of individual

[1] https://github.com/emorynlp/elit

speakers is important especially in multi-party interaction. In comparison to the binary indicator used in BL and SR that can handle only up to two speakers, adding the subtask for speaker-grounded encoding is shown to perform better for multi-speaker dialogues: the improvement of OR+SG over SR is 3.5% F1 for *Friends*, but around 1% F1 for the other two. Our statistics show that 43% dialogues in *Friends* have at least 4 speakers, while being only 15% and 24% for the other two, suggesting that the multi-speaker environment indeed benefits more from the new speaker encoding scheme.

In addition, the percentages of pronouns in the gold mentions are 80.3%, 53.5%, and 63.5% in *Friends*, *Onto-Conv*, and *BOLT* respectively, which also highlights the importance of a better encoding scheme to handle a large portion of pronouns present in dialogue. Thus, we suggest to employ a more advanced dialogue encoding that utilizes the speaker interaction clues as one of the future research direction for this online-decoding task.

4 Conclusion

This paper presents a new coreference resolution direction that aims towards an online decoding setting for dialogue processing. A baseline and four incremental-updated models are proposed and evaluated on three datasets of the dialogue domain, and the best-performing model shows significant improvement over the baseline by ≈10% F1. Further analysis suggests the importance of mention recall and speaker encoding, which could serve as the next future directions of this online setting.

References

Kevin Clark and Christopher D. Manning. 2015. Entity-centric coreference resolution with model stacking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1405–1415, Beijing, China. Association for Computational Linguistics.

Kevin Clark and Christopher D. Manning. 2016. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Berlin, Germany. Association for Computational Linguistics.

Yifan Gao, Piji Li, Irwin King, and Michael R. Lyu. 2019. Interconnected question generation with coreference alignment and conversation flow modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4853–4862, Florence, Italy. Association for Computational Linguistics.

Han He, Liyan Xu, and Jinho D. Choi. 2021. Elit: Emory language and information toolkit.

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.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Xuansong Li, Martha Palmer, Nianwen Xue, Lance Ramshaw, Mohamed Maamouri, Ann Bies, Kathryn Conger, Stephen Grimes, and Stephanie Strassel. 2016. Large multi-lingual, multi-level and multi-genre annotation corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 906–913, Portorož, Slovenia. European Language Resources Association (ELRA).

Yangming Li, lemao liu, and Shuming Shi. 2021. Empirical analysis of unlabeled entity problem in named entity recognition. In *International Conference on Learning Representations*.

Fei Liu, Luke Zettlemoyer, and Jacob Eisenstein. 2019. The referential reader: A recurrent entity network for anaphora resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5918–5925, Florence, Italy. Association for Computational Linguistics.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph LSTMs. *Transactions of the Association for Computational Linguistics*, 5:101–115.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Jun Quan, Deyi Xiong, Bonnie Webber, and Changjian Hu. 2019. GECOR: An end-to-end generative ellipsis and co-reference resolution model for task-oriented dialogue. 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 4547–4557, Hong Kong, China. Association for Computational Linguistics.

Sunil Kumar Sahu, Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Inter-sentence relation extraction with document-level graph convolutional neural network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4309–4316, Florence, Italy. Association for Computational Linguistics.

Shubham Toshniwal, Sam Wiseman, Allyson Ettinger, Karen Livescu, and Kevin Gimpel. 2020. Learning to Ignore: Long Document Coreference with Bounded Memory Neural Networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8519–8526, Online. Association for Computational Linguistics.

Bo-Hsiang Tseng, Shruti Bhargava, Jiarui Lu, Joel Ruben Antony Moniz, Dhivya Piraviperumal, Lin Li, and Hong Yu. 2021. CREAD: Combined resolution of ellipses and anaphora in dialogues. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3390–3406, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Sam Wiseman, Alexander M. Rush, Stuart Shieber, and Jason Weston. 2015. Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1416–1426, Beijing, China. Association for Computational Linguistics.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online. Association for Computational Linguistics.

Patrick Xia, João Sedoc, and Benjamin Van Durme. 2020. Incremental neural coreference resolution in constant memory. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8617–8624, Online. Association for Computational Linguistics.

Liyan Xu and Jinho D. Choi. 2020. Revealing the myth of higher-order inference in coreference resolution. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8527–8533, Online. Association for Computational Linguistics.

Liyan Xu and Jinho D. Choi. 2021. Adapted end-to-end coreference resolution system for anaphoric identities in dialogues. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 55–62, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ethan Zhou and Jinho D. Choi. 2018. They exist! introducing plural mentions to coreference resolution and entity linking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 24–34, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

A Appendix

A.1 Dataset

The annotation in *Friends* includes plural links where a mention can belong to more than one entity clusters. We discard those mentions with plural links in our experiments and leave them as future work. All remaining mentions for *Friends* are personal mentions.

BOLT does not come with a predefined train/dev/test split. We use a random split of 80%, 10%, 10% of documents in each genre for the train/dev/test split. In addition, we only use genres "en" and "sm" in *BOLT*, as other genres currently do not have user IDs provided and only constitute less than 5% documents of entire dataset. The details of our split are provided in `https://github.com/lxucs/online-bolt`.

A.2 Implementation

For training on entire dialogue contexts as document input (`BL` and `SR`), we follow the similar hyperparameter settings as Joshi et al. (2019, 2020); Xu and Choi (2020), where long documents are split into independent segments with the maximum sequence length of 384 for $\text{SpanBERT}_{\text{BASE}}$. We employ the learning rate of 2×10^{-5} for BERT parameters and 2×10^{-4} for task parameters with the dropout rate as 0.3. Maximum span length is set to 6 for *Friends* and 25 for *Onto-Conv* and *BOLT*. In the coarse pruning stage, we keep a maximum number of antecedents as 20 for *Friends* and 50 for *Onto-Conv* and *BOLT*.

For online training and inference on the utterance sequence input (`OR`, `+SG`, `+SA`), we use one BERT segment so that the length of current utterance with past context does not exceed 384 tokens in our experiments. Gradient accumulation of 16 steps is applied during online training. We use the same learning rates and training epochs, similar as training on document input. Our best model has $\alpha_c = 1, \alpha_m = 0.1, \alpha_s = 0.1$ for the multi-task learning.

All experiments are conducted on NVIDIA TITAN RTX GPUs with 24GB memory. Training on document input takes around 3 hours and training on online input takes around 8-12 hours. All proposed methods have similar inference time, as they follow similar architecture and all operate on the online inference for prediction.