Pipeline Coreference Resolution Model for Anaphoric Identity in Dialogues

Damrin Kim*, **Seongsik Park**, **Mirae Han** and **Harksoo Kim**
Department of Artificial Intelligence, Konkuk University, Seoul, Republic of Korea
{ekafls33, a163912, future26, nlpdrkim}@konkuk.ac.kr

Abstract

CODI-CRAC 2022 Shared Task in Dialogues consists of three sub-tasks: Sub-task 1 is the resolution of anaphoric identity, sub-task 2 is the resolution of bridging references, and sub-task 3 is the resolution of discourse deixis/abstract anaphora. Anaphora resolution is the task of detecting mentions from input documents and clustering the mentions of the same entity. The end-to-end model proceeds with the pruning of the candidate mention, and the pruning has the possibility of removing the correct mention. Also, the end-to-end anaphora resolution model has high model complexity, which takes a long time to train. Therefore, we proceed with the anaphora resolution as a two-stage pipeline model. In the first mention detection step, the score of the candidate word span is calculated, and the mention is predicted without pruning. In the second anaphora resolution step, the pair of mentions of the anaphora resolution relationship is predicted using the mentions predicted in the mention detection step. We propose a two-stage anaphora resolution pipeline model that reduces model complexity and training time, and maintains similar performance to end-to-end models. As a result of the experiment, the anaphora resolution showed a performance of 68.27% in Light, 48.87% in AMI, 69.06% in Persuasion, and 60.99% on Switchboard. Our final system ranked 3rd on the leaderboard of sub-task 1.

1 Introduction

Anaphora resolution(Kim et al., 2021; Yu et al., 2022) is the task of detecting mentions from input documents and clustering the mentions of the same entity. It is used for various natural language processing tasks such as document summarization, question answering, and knowledge extraction. Mention detection is the task of extracting candidate word spans that are likely to be mentions within a sentence. Mention refers to a span of candidate words that are highly likely to have a anaphora relationship in a sentence. Most of the anaphora resolution models being studied recently are end-to-end models. The end-to-end model extracts and learns all candidate word spans that are likely to be a mention and prunes them at a fixed ratio. The mention pairs are made from pruned mentions and are clustered into final mention pairs based on calculated scores. However, fixing the prune ratio is inefficient. A high pruning ratio increases the number of non-correct candidate mentions, increasing the amount and complexity of calculations. Conversely, a low ratio increases the possibility of removing correct answers instead of lowering the amount and complexity. Finding the optimal pruning ratio is important because the pruning ratio of the mention detection can directly affect the anaphora resolution performance. Therefore, we propose a two-stage anaphora resolution pipeline model to speed up training and reduce model complexity without pruning. Table 1 summarizes the description of the system and experiment.

In the first mention detection step, the mention is trained by calculating scores of all possible candidate word spans in the input sentence. In the second anaphora resolution step, a mention pair consists of the mentions predicted in the detection step. Then, the mention pair score is calculated to train the mention pair, which is a anaphora relationship. The proposed model shows high performance in the mention detection. Moreover, compared with the self-reimplemented end-to-end anaphora resolution model, it shows similar performance and fast training speed.

2 Related Works

Recently, anaphora resolution has been studied using an end-to-end model that learns pairwise scores of entity mentions(Lee et al., 2017). The end-to-end model calculate mention score with all possible spans in a given text. The pruning step proceeds

Proceedings of the CODI-CRAC 2022 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue, pages 28–31
Gyeongju, Republic of Korea, October 17, 2022.

Track	Resolution of anaphoric identities
Setting	Predicted mentions
Baseline	-
Approach	Sec. 3.1 and 3.2
Train Data	Sec. 4.1
Dev Data	Sec. 4.1

Table 1: System summary

with the calculated mention scores. The anaphora score is calculated by a pair of mentions made with current and antecedent mentions(Lee et al., 2018; Devlin et al., 2018; Joshi et al., 2020).

Before Dobrovolskii (2021) was introduced, the end-to-end models mainly achieved a state-of-the-art anaphora resolution. Dobrovolskii (2021) proceeded with a pipeline to resolve anaphora resolution. As a result, they reduced the complexity of the model from $O(n^4)$ to $O(n^2)$ and improved its performance. Unlike the existing end-to-end models, it is possible to efficiently detect mentions because it does not calculate mention scores and perform the pruning step. We propose a two-stage anaphora resolution model that utilizes not only the information of the current speaker but also of the previous speaker, considering the anaphora resolution characteristics of the dialogue domain. The proposed model is faster in training and evaluation compared to end-to-end models.

3 Model

3.1 Mention Detection

The mention detection model consists of a pre-trained language model, a mention representation generation layer, and a mention score generation layer.

$$X = \{x_1, x_2, \cdots, x_T\} \tag{1}$$

Pre-trained language model receives input tokens in a sentence and outputs the token representation X. T denotes the number of tokens. $N = T(T+1)/2$ is the number of possible text spans.

$$g_m(i) = [x_{START(i)}, x_{END(i)}] \tag{2}$$

Mention representation $g_m(i)$ is generated by connecting $START(i)$ and $END(i)$, which are the start and end index token representations of span i. The mention score $S_m(i)$ is calculated through FNN(feed-forward neural network):

$$S_m(i) = W_m \cdot FNN_m(g_m(i)) \tag{3}$$

$S_m(i)$ is calculated by multiplying the mention representation by the learnable weight W_m. It trains to minimize the cross-entropy between predicted and correct mentions, as follows:

$$loss_m = -\sum_i Y_i^m log\left(\hat{Y}_i^m\right) \tag{4}$$

3.2 Anaphora Resolution

Anaphora resolution model can be divided into a pre-trained language model, a mention representation generation layer, and a pairwise score generation layer. The pre-trained language model receives input tokens in a document and outputs the token representation X. D denotes the number of tokens in the document. We segment a document into the maximum size of pre-trained language model to process documents that are longer than this. The segmented documents are used independently as input. The outputs of the pre-trained language model are concatenated and reconstructed to be a document.

$$X = \{x_1, x_2, ..., x_D\} \tag{5}$$

Mention representation $g_c(i)$ is generated using the predicted mentions in the mention detection model. The token representations of span boundaries, the average of token representations in span, and the feature vector are concatenated to generate $g_c(i)$. The feature vector $\phi(i)$ contains speaker information of current and previous sentences and is initialized by random embedding. This helps eliminate the ambiguity of personal pronouns such as 'you' and 'I' when there are multiple speakers.

$$\begin{aligned} g_c(i) = [&x_{START(i)}, x_{END(i)} \\ &, avg(x_{START(i)}; x_{END(i)}), \phi(i)] \end{aligned} \tag{6}$$

Mention pair uses mention representations to generate all possible pairs without duplicate ones. Next, pairwise score $S_c(i,j)$ is calculated through FNN by connecting $g_c(i)$ and $g_c(j)$, which are mention representation pairs:

$$S_c(i,j) = W_c \cdot FNN_c(g_c(i), g_c(j)) \tag{7}$$

$S_c(i)$ is calculated by multiplying the mention representation pair by the learnable weight W_c. It trains to minimize the cross-entropy between predicted pairwise scores of mention pairs and correct mention pairs:

$$loss_c = -\sum_i Y_i^c log\left(\hat{Y}_i^c\right) \tag{8}$$

4 Experiments

4.1 Datasets

We use datasets provided by CODICRAC 2022 Shared-Task for learning and evaluation. We use the train and dev dataset of Light, AMI, Persuasion, Switchboard and train, dev, and test dataset of ARRAU for training, and use the test dataset of Light, AMI, Persuasion, and Switchboard for evaluation. All datasets are dialogue domains and consist of Universal Anaphora(Poesio et al., 2004) annotations. The statistics of the datasets used for training and validation are shown in Table 2 and 3. #D is the total number of documents, #S is the total number of sentences, #W is the total number of words, #M is the total number of mentions, #C is the total number of clusters, and #SPK is the average number of speakers per document.

	Light	AMI	PSUA	SWBD	ARRAU
#D	20	7	21	11	202
#S	909	4,140	813	1,343	4,230
#W	11,495	33,741	9,185	14,992	110,440
#M	3,907	8,918	2,743	4,024	34,454
#C	1,803	4,391	1,513	2,362	23,238
#SPK	2,95	4	2	2	-

Table 2: Statistics for train datasets.

	Light	AMI	PSUA	SWBD	ARRAU
#D	21	3	27	22	18
#S	924	1,968	1,110	3,653	479
#W	11,824	18,260	12,198	35,027	12,845
#M	3,941	4,870	3,697	9,392	3,961
#C	1,789	2,551	1,996	5,436	2,640
#SPK	3	4	2	2	-

Table 3: Statistics for dev datasets.

4.2 Evaluation Metrics

The Mention Detection Model measures performance using F1-score, the harmonic mean of precision and recall, as follows:

$$
\begin{aligned}
Precision &= \frac{True\ Positive}{True\ Positive + False\ Positive} \\
Recall &= \frac{True\ Positive}{True\ Positive + False\ Negative} \\
F1 - score &= \frac{2 * Precision * Recall}{Precision + Recall}
\end{aligned} \tag{9}
$$

The evaluation of the anaphora resolution model is conducted with the SemEval evaluation program. We measure CoNLL F1 score(Pradhan et al., 2014) which averages three performances in the official evaluation process since CoNLL-2011: B3(Bagga and Baldwin, 1998), a mention-based method, CEAFe(Luo, 2005), an entity-based method and MUC(Vilain et al., 1995), a link-based method.

4.3 Experiments on Mention Detection

As shown in Table 4, our mention detection model shows F1 performance of 92.17% on Light, 80.46% on AMI, 89.67% on Persuasion(PSUA), and 85.02% on Switchboard(SWBD).

	Precision	Recall	F1-score
Light	94.76	89.72	92.17
AMI	88.15	74.01	80.46
PSUA	90.67	88.70	89.67
SWBD	92.60	78.58	85.02

Table 4: Results on mention detection for test datasets.

4.4 Experiments on Anaphora Resolution

As shown in Table 5, our anaphora resolution model shows a CoNLL F1 performance of 68.27% on Light, 48.87% on AMI, 69.06% on Persuasion, and 60.99% on Switchboard.

		Light	AMI	PUSA	SWBD
MUC	P	73.45	36.05	70.04	53.83
	R	83.31	77.67	83.23	83.12
	F1	78.07	49.24	76.07	65.34
B^3	P	76.72	46.22	70.00	58.46
	R	55.14	64.06	69.97	69.08
	F1	64.16	53.70	69.99	63.33
$CEAF_e$	P	63.08	70.76	76.31	70.73
	R	62.07	31.57	51.00	44.07
	F1	62.27	43.66	61.14	54.31
CoNLL F1	F1	68.27	48.87	69.06	60.99

Table 5: Results on anaphora resolution for test datasets.

In Table 6, the proposed model shows similar performance to the self-implemented end-to-end anaphora resolution model(Lee et al., 2017).

We also show the effectiveness of the two-stage pipeline model because the model complexity is reduced from $O(n^4)$ to $O(n^2)$, and the total training time is reduced by about 1/10.

model	Light	AMI	PSUA	SWBD
end-to-end	70.45	35.34	67.52	61.27
ours	68.27	48.87	69.06	60.99

Table 6: CoNLL F1-score of pipeline(proposed model) and end-to-end model

5 Conclusion

We propose a pipeline model for anaphora resolution. Our proposed model consists of a mention detection model and an anaphora resolution model. The mention detection model predicts mentions by the span prediction method. The anaphora resolution model predicts a pair of mentions of an anaphora relation by the mention pair method based on results from the mention detection model. In subtask 1, our model achieved 68.3%, 48.8%, 69.1%, and 61.0% performance on Light, AMI, Persuasion, and Switchboard (ranked in the top 3). We will study a mention detection model robust in noun phrases by reflecting the context of the document and an anaphora resolution model by using GNN to reflect structural information between mentions.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(No.2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques)

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566. Citeseer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Vladimir Dobrovolskii. 2021. Word-level coreference resolution. *arXiv preprint arXiv:2109.04127*.

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.

Hongjin Kim, Damrin Kim, and Harksoo Kim. 2021. The pipeline model for resolution of anaphoric reference and resolution of entity reference. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 43–47.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. *arXiv preprint arXiv:1804.05392*.

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.

Massimo Poesio, Rahul Mehta, Axel Maroudas, and Janet Hitzeman. 2004. Learning to resolve bridging references. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 143–150.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2014, page 30. NIH Public Access.

Marc Vilain, John D 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*.

Juntao Yu, Sopan Khosla, Ramesh Manuvinakurike, Vincent Ng, Massimo Poesio, Michael Strube, and Carolyn Rosé. 2022. The codi-crac 2022 shared task on anaphora, bridging, and discourse deixis in dialogue. In *Proceedings of the CODI-CRAC 2022 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*.