Uniform Anaphora Resolution and Evaluation

Anonymous ACL submission

Abstract

Despite the increasing attention on tackling anaphora resolution in an end-to-end multitask learning fashion, the state of the research topic is still unsatisfactory in that most works focus only on a subset of relations (either bridging or coreference), lacking generalizability and granularity for more complicated anaphoric relations. Moreover, the evaluations are still a mix of diverse metrics for different subtasks. We leverage a multitask learning framework from the Relation Extraction field which can be extended to perform fine-grained anaphora resolution and introduce a heterogeneous graph representation to evaluate coreference and other anaphoric relations using one uniform metric. All the data and source code will be publicly available. [1]

1 Introduction

Anaphora resolution is the task of linking nominal expressions to entities in the context of interpretation [2]. Typical anaphoric relations include coreference, where two mentions refer to the same entity, as well as bridging reference, which indicates an associative, non-coreferential relation.

Although recent works, such as PairSpanBERT (Kobayashi et al., 2023) enable anaphora resolution in a 'real-world' setting, meaning that given a raw document, the system can predict the coreferent or bridging anaphor and their antecedents, there are still a few weaknesses to the current methods. First, the anaphoric relations are predefined and lack generalizability. The models do not allow extensions for other task-specific associative relations. Second, previous works can only predict at a coarse level, treating bridging relations as one anaphoric relation even though the annotation has a finer granularity. Third, although the training is done jointly with all types of the relations, different evaluation metrics are used for different types. The existing metrics are originally designed for subtasks of anaphora resolution, and it is very hard to evaluate globally across different anaphoric types. To address the aforementioned issues, we trained a fine-level multi-class joint model for both problems, by adopting a multitask learning framework, an approach to solve the relation extraction problem. We also propose a metric that can be uniformally applied for evaluating both types of anaphora resolution problems.

2 Related Work

End-to-end coreference resolution models (Lee et al., 2017, 2018; Xu and Choi, 2020; Wu et al., 2020; Kirstain et al., 2021) refer to systems that can detect candidate mentions and find their possible antecedents. In contrast to studies on applicability of such end-to-end framework to the coreference resolution problem, a broader problem of bridging resolution is less studied and the vast majority of existing bridging resolution systems are evaluated in rather unrealistic settings, where the gold mentions are assumed as input (Hou et al., 2014; Roesiger et al., 2018; Hou et al., 2018; Yu and Poesio, 2020).

Kobayashi et al. (2022b) is the first attempt to evaluate bridging resolution in an end-to-end setting where a resolver needs to identify bridging relations given a raw document. It is also found from previous works (Yu and Poesio, 2020) that training coreference and bridging resolvers jointly is beneficial to both tasks. As a result, more research in document level end-to-end multitask learning anaphora resolution has recently emerged (Kobayashi et al., 2022b,a, 2023). Although the models are jointly trained, only Kobayashi et al. (2022b) evaluates the coreference resolver of their framework. The others only report the performances of the bridging resolvers. Moreover, the experiments in Kobayashi

[1] Anonymized for reviewing.

[2] We do not consider event anaphora (Sukthanker et al., 2020; Xie et al., 2023) for the present paper.

et al. (2022b) are conducted on ISNotes (Markert et al., 2012) and BASHI (Rösiger, 2018). To our best knowledge, this paper is the first attempt to uniformly evaluate different subtypes of anaphoric resolution systems, as well as the first work to use the ARRAU corpus for evaluating anaphora resolvers.

There has been a growing body of literature that explores anaphoric relations tailored to task-oriented text, considering the diverse natures of the corpora involved. (Fang et al., 2021, 2022; Rim et al., 2023). However, due to their vastly different contextual characteristics, we argue that it is necessary to distinguish the anaphoric relations found in instruction-heavy text. We propose to refer to bridging as *static anaphoric relations*, where the definition is fixed and refers to the lexical and referential bridging defined by Clark (1975). Task-specific relations are called *dynamic*, as objects in these relations tend to change according to the types of texts being annotated. Besides the capability to predict static anaphoric relations, our model also tries to generalize on the dynamic relations.

Traditionally, MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin), and CEAF (Luo, 2005) have been popular choices as evaluation metrics for coreference resolution, where the metrics emphasize link, mention, and entity respectively. An average score of the three metrics is often reported as a comprehensive evaluation result. All these metrics are mathematically clean and elegant. However, the fact that these metrics are designed for coreference alone makes it difficult to generalize to non-identity anaphoric relations. Specifically, link-based metrics are designed for identity links and can only handle one relation type. Moreover, mention-entity based metrics assume unique set-membership where each entity can only have one membership at a time. This constraint renders the evaluation of split-antecedents impossible without creating intermediate accommodation sets. It also prohibits the evaluation of mentions existing in different clusters based on different relations. But in reality, a mention in a coreference cluster can sometimes hold other anaphoric relation types (e.g., bridging relation) with another mention.

When it comes to bridging resolution, intuitively, people use pairwise F1 as the metric since bridging relations are annotated in pairs. Despite the straightforwardness, this metric lacks the capability to propagate the relation from one mention to other mentions that are not immediately connected to it. Our proposed uniform metric can handle all these challenges while also evaluating the system globally.

3 Document-level Anaphora Graph

We propose a graph representation as a unified output representation specifically for end-to-end anaphora resolution frameworks, where all three subtasks can be evaluated simultaneously. Furthermore, a graph data structure can help address the current issues with evaluation discussed above. Formally, given a document D, we convert it into a graph G by first adding a document vertex v_D. For all the mentions in the document $\{m_i\}_{i=1}^{M}$, they are turned into vertices $\{v_i\}_{i=1}^{M}$. We connect v_D with all mentions with edges e_D. This is for evaluation of mention extraction. For mentions in a coreference cluster c_j, their corresponding vertices are fully connected by coreference edges e_j. If mention m_{a1} in coreference cluster c_a and mention m_{b1} in cluster c_b hold a bridging relation r, then vertices in c_a and vertices in c_b are fully connected by that relation edge e_r. We use fully connected graphs instead of minimum spanning to properly penalize the system missing a bridging edge between two mentions if any of which belongs to a large cluster, which is usually more informative and critical to understand text. This heterogeneous graph structure allows for evaluation at a fine level as well as the extension of dynamic relations. And it also addresses the inability of previous evaluation metrics to handle split-antecedents. While we sympathize with the method of accommodation sets (Paun et al., 2022), we find it semantically more natural to treat split-antecedents as a bridging relation rather than coreference. Refer to A.1 to see more subgraph examples.

4 Datasets

We select ARRAU RST (Poesio and Artstein, 2008; Uryupina et al., 2020) and CUTL (Rim et al., 2023) as our two datasets. ARRAU RST contains annotations of coreference and static anaphoric relations on newswire texts. CUTL is annotated on recipe data with coreference and dynamic anaphoric relations such as Transformation where an ingredient has undergone some transformations, e.g., baking, but remain substance identical (Ye et al., 2023).

5 Baseline System

We adapt the relation extraction model TAG from Zhang et al. (2023) and use it as the baseline for our experiments. It is an end-to-end joint extraction model of entities and relations trained and evaluated on the DOCRED dataset (Yao et al., 2019).

Given a raw document, the extraction process consists of three subtasks: (1) Mention extraction; (2) Coreference resolution; and (3) Relation extraction. The mention extraction task extracts all possible spans for entities formulated as a tagging task. Both coreference resolution and relation extraction are trained jointly with a table-to-graph generation model.

For our task, we keep the mention extractor and coreference resolver as they are and formulate bridging resolution as an extraction task of anaphoric relations by replacing the predefined relations from DOCRED to corresponding bridging references in our anaphora corpora. To alleviate the problem of data imbalance between coreference and bridging examples, we adapt the model by adding a new hyperparameter, β, to balance coreference and bridging relation loss. The new loss function is defined as follows, where $\mathcal{L}_{tc}$ and $\mathcal{L}_{gc}$ are coarse and fine level coreference extraction losses, $\mathcal{L}_{tr}$ and $\mathcal{L}_{gr}$ are bridging extraction losses, and α is a hyperparameter balancing coarse and fine level loss.

$$\begin{aligned}\mathcal{L} &= \beta \cdot \mathcal{L}_{tc} + \mathcal{L}_{tr} \\ &+ \alpha \cdot (\beta \cdot \mathcal{L}_{gc} + \mathcal{L}_{gr})\end{aligned} \tag{1}$$

6 Experiments

Model training and parameter tunining We train and evaluate the TAG model on ARRAU and CUTL datasets. We use the Roberta-base model (Liu et al., 2019) as the encoder and reuse the hyperparameters from Zhang et al. (2023). We train for 100 epochs on the mention extraction task and 200 epochs for coreference and bridging resolution. We set the balancing loss weight β to 0.3 after fine tuning on the ARRAU dev set. The training process takes about 3 hours on a TITAN Xp GPU.

Data preprocessing Given the high complexity of the model and the limitation of our computation resources, we are forced to filter out documents of over 400 tokens to avoid out-of-memory problem.

And to recover all the 'drop arguments' in CUTL documents, we follow their Dense Paraphrasing pipeline, specifically using GPT3 to paraphrase the missing arguments in the surface form. Note that the paraphrasing would generate new tokens in the text, so this step happens before filtering.

Train-test partition Table 1 shows the statistics and train-test distribution of the two datasets after filtering. We follow the train-test split of the original datasets.

	Train	Dev	Test	All
ARRAU RST	104	5	14	123
CUTL	80	N/A	20	100

Table 1: Train-test split of filtered data. CUTL did not release a separate 'dev' split

	P	R	F1
ARRAU	83.29	65.68	73.44
CUTL	92.92	95.37	94.13

Table 2: Results of mention extraction.

7 Results

We evaluate the performance of our baseline model in two settings: (1) THE LOCAL LEVEL: Since the end-to-end anaphoric resolution consists of three subtasks, we report the individual performance on each task using the standard metrics for them. (2) THE GLOBAL LEVEL: We convert the extracted mentions and anaphoric relations into a heterogeneous graph and use a triple-based measure as a unified metric.

7.1 Local Level Evaluation

Mention extraction Results of mention extraction are reported in table 2, where precision, recall and F1 scores are calculated on the entire set of mentions where the mention can be either a direct anaphor or a bridging anaphor. The F1 scores on both datasets are very good, especially for CUTL where the score is over 0.94.

Coreference resolution For coreference resolution, we use the standard MUC, B^3 and CEAF metrics and their unweighted average CoNLL score. Table 3 shows the results on both datasets. The TAG model achieves an average CoNLL score of 0.45 on ARRAU. This is lower than other coreference specific models like Yu et al. (2020), indicating that ARRAU is a challenging corpus that needs extra features and tuning to achieve good performance. As for the results on CUTL data, the TAG model achieves an average CoNLL score of 0.85. To invesigate the competence of our baseline model, we run the model from Rim et al. (2023)

		MUC	B^3	$CEAF_e$	CoNLL
ARRAU	TAG	61.06	59.07	14.28	44.80
CUTL	Rim et al. (2023)	85.22	40.03	45.72	57.00
	TAG	89.11	94.91	69.76	84.59

Table 3: Results of coreference resolution on ARRAU and CUTL. We also report the score of Rim et al. (2023) on CUTL.

	P	R	F1
PairSpanBERT	21.2	16.9	18.8
TAG	16.67	4.65	7.27

Table 4: Results of coarse level bridging resolution on ARRAU RST.

Relation	P	R	F1
Other	50.00	5.26	9.52
Split-antecedents	66.67	8.33	14.81

Table 5: Results of fine level bridging resolution on ARRAU RST.

on CUTL and the results show that the TAG-based model achieves significantly better F1 score.

Bridging resolution We use precision, recall and F1 as the standard metrics for bridging resolution. Table 4 shows the results of our baseline model at coarse level where all bridging relations are collapsed into one. PairSpanBERT is also doing bridging resolution at a coarse level in a very similar end-to-end setting, so we include their reported scores here. It is worth noting that this is not fully comparable since we used a subset of ARRAU and the PairSpanBERT implementation is not released, which makes it hard to replicate their model in our setting. However, the higher scores suggest that TAG is also not performing well on bridging resolution on newswire texts. We also report the results of a fine-grained bridging resolution in table 5. We only include two relations, i.e., other and split-antecedents, in the table because the model can only predict these two relations correctly. The comparatively higher scores on split-antecedents imply that the model can understand the aggregation of singular mentions to plural references. Nevertheless, the low overall performance indicates that bridging resolution at a fine level is still very challenging. Finally, the results of a fine level bridging resolution on CUTL data are presented in table 6. Except for the two relations that both models failing to predict correctly, i.e., Metonym and Separation, the TAG model is also unable to predict any Meronym relation but achieves better F1 scores on

	Rim et al. (2023)			TAG		
Relation	P	R	F1	P	R	F1
MERONYM	25.57	7.11	11.13	N/A	N/A	N/A
TRANS.	82.51	58.21	68.26	81.20	86.75	83.88
AGG.	82.03	59.23	68.79	82.18	69.75	75.46

Table 6: Results of fine level bridging resolution on CUTL.

		P	R	F1
ARRAU	TAG	67.00	46.26	54.73
CUTL	Rim et al. (2023)	69.25	60.13	64.37
	TAG	84.74	79.14	81.84

Table 7: Smatch score of TAG on ARRAU RST and CUTL. We also report the score of Rim et al. (2023) on CUTL.

Transformation and Aggregation. The failure to perform prediction is most likely due to the small number of examples for those relations. The result suggests that TAG generalizes well on dynamic anaphoric relations.

7.2 Global Level Evaluation

We use the document-level anaphora graph we propose as the data structure and use a triple-based metric inspired by Smatch (Cai and Knight, 2013) for evaluation. The score is computed in terms of the matching triples (v_i, e_j, v_k) in the graphs using precision, recall and F1. Given two graphs G_{pred} and G_{gold}, m is the number of matching triples, t is the total number of triples in the first graph, g is the total number of triples in the second graph. M, T and G are the sum of m, t and g of all documents in the evaluation batch. The precision score is defined as $P = M/T$, recall is defined as $R = M/G$. The final Smatch score (F) is their harmonic mean. Table 7 shows the results of our model on the two datasets as well as our replication of the model from Rim et al. (2023) on CUTL. It can be seen that the model achieves higher scores on CUTL than ARRAU showing that anaphora resolution on ARRAU is a more challenging task, which can be attributed to the complexity of newswire texts and their annotation scheme. Also, TAG outperforms the CUTL model by a large margin, which aligns with the standard metrics in our previous experiments. This suggests that TAG is an overall more competent model on anaphora resolution.

8 Conclusion

We adapt an end-to-end relation extraction framework into the field of anaphora resolution and show that the model is competent on extracting dynamic anaphoric relations. We also propose a graph based metric to evaluate anaphoric resolution system in an end-to-end setting. This generalized method can accommodate dynamic anaphoric relations and evaluate the system at the global level.

9 Limitations

We train and evaluate models for end-to-end anaphora resolution at a fine level and propose a new uniform metric. Given the fact that there is no related work that conducts experiments in the same setting as we do, plus we only use a subset of data, our results on ARRAU RST are not fully comparable to any previous work. In addition, since we exclude documents of long sequence, further investigations are needed to evaluate the model on long documents where complex anaphoric relations may occur more frequently.

References

Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chains. Citeseer.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria. Association for Computational Linguistics.

Herbert H. Clark. 1975. Bridging. In *Theoretical Issues in Natural Language Processing*.

Biaoyan Fang, Timothy Baldwin, and Karin Verspoor. 2022. What does it take to bake a cake? the RecipeRef corpus and anaphora resolution in procedural text. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3481–3495, Dublin, Ireland. Association for Computational Linguistics.

Biaoyan Fang, Christian Druckenbrodt, Saber A Akhondi, Jiayuan He, Timothy Baldwin, and Karin Verspoor. 2021. ChEMU-ref: A corpus for modeling anaphora resolution in the chemical domain. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1362–1375, Online. Association for Computational Linguistics.

Yufang Hou, Katja Markert, and Michael Strube. 2014. A rule-based system for unrestricted bridging resolution: Recognizing bridging anaphora and finding links to antecedents. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2082–2093, Doha, Qatar. Association for Computational Linguistics.

Yufang Hou, Katja Markert, and Michael Strube. 2018. Unrestricted bridging resolution. *Computational Linguistics*, 44(2):237–284.

Yuval Kirstain, Ori Ram, and Omer Levy. 2021. Coreference resolution without span representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 14–19, Online. Association for Computational Linguistics.

Hideo Kobayashi, Yufang Hou, and Vincent Ng. 2022a. Constrained multi-task learning for bridging resolution. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 759–770, Dublin, Ireland. Association for Computational Linguistics.

Hideo Kobayashi, Yufang Hou, and Vincent Ng. 2022b. End-to-end neural bridging resolution. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 766–778, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Hideo Kobayashi, Yufang Hou, and Vincent Ng. 2023. PairSpanBERT: An enhanced language model for bridging resolution. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6931–6946, Toronto, Canada. 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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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.

Katja Markert, Yufang Hou, and Michael Strube. 2012. Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 795–804, Jeju Island, Korea. Association for Computational Linguistics.

Silviu Paun, Juntao Yu, Nafise Sadat Moosavi, and Massimo Poesio. 2022. Scoring coreference chains with split-antecedent anaphors.

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).

Kyeongmin Rim, Jingxuan Tu, Bingyang Ye, Marc Verhagen, Eben Holderness, and James Pustejovsky. 2023. The coreference under transformation labeling dataset: Entity tracking in procedural texts using event models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12448–12460, Toronto, Canada. Association for Computational Linguistics.

Ina Roesiger, Arndt Riester, and Jonas Kuhn. 2018. Bridging resolution: Task definition, corpus resources and rule-based experiments. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3516–3528, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ina Rösiger. 2018. BASHI: A corpus of Wall Street Journal articles annotated with bridging links. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. 2020. Anaphora and coreference resolution: A review. *Information Fusion*, 59:139–162.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa J. Rodriguez, and Massimo Poesio. 2020. Annotating a broad range of anaphoric phenomena, in a variety of genres: the arrau corpus. *Natural Language Engineering*, 26(1):95–128.

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*.

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.

Dongdong Xie, Fei Li, Bobo Li, Chong Teng, Donghong Ji, and Meishan Zhang. 2023. Chinese event discourse deixis resolution: Design of the dataset and model. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(11):1–26.

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.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Bingyang Ye, Jingxuan Tu, and James Pustejovsky. 2023. Scalar anaphora: Annotating degrees of coreference in text. In *Proceedings of The Sixth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2023)*, pages 28–38.

Juntao Yu and Massimo Poesio. 2020. Multitask learning-based neural bridging reference resolution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3534–3546, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Juntao Yu, Alexandra Uma, and Massimo Poesio. 2020. A cluster ranking model for full anaphora resolution. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 11–20, Marseille, France. European Language Resources Association.

Ruoyu Zhang, Yanzeng Li, and Lei Zou. 2023. A novel table-to-graph generation approach for document-level joint entity and relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10853–10865, Toronto, Canada. Association for Computational Linguistics.

A Appendix

A.1 Graph Representation Examples

We present two graph representations of the anaphoric relations in documents in ARRAU and CUTL in figure 1 and 2. For the purpose of better visualization, we only show subgraphs of them. The value of the vertex is shown in the form of {mention $sentence_{index}.word_{index}$}. The dotted edge refers to the document edge. Edges of different colors refer to the relations of matching color.

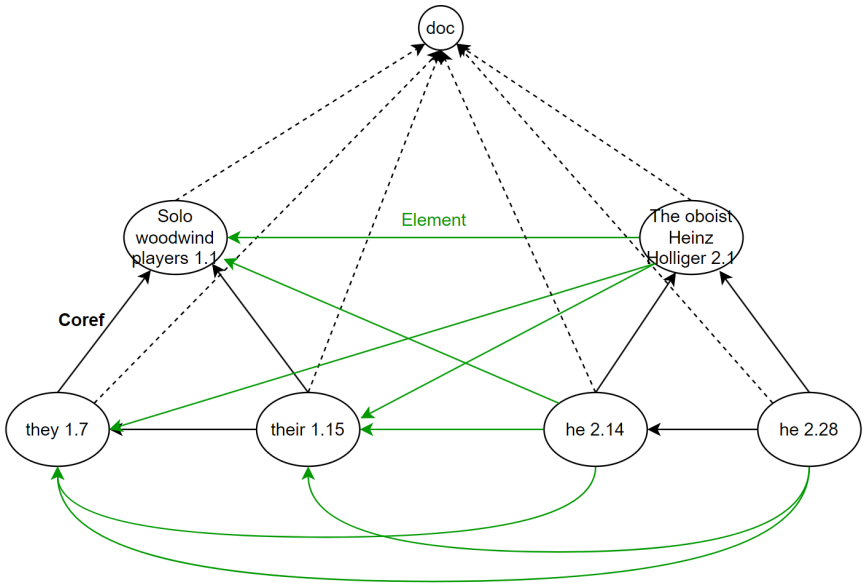

Figure 1: Graph representation of anaphoric relations in a document in ARRAU. The black edges are coreference relations. The green edges refer to Element relation.

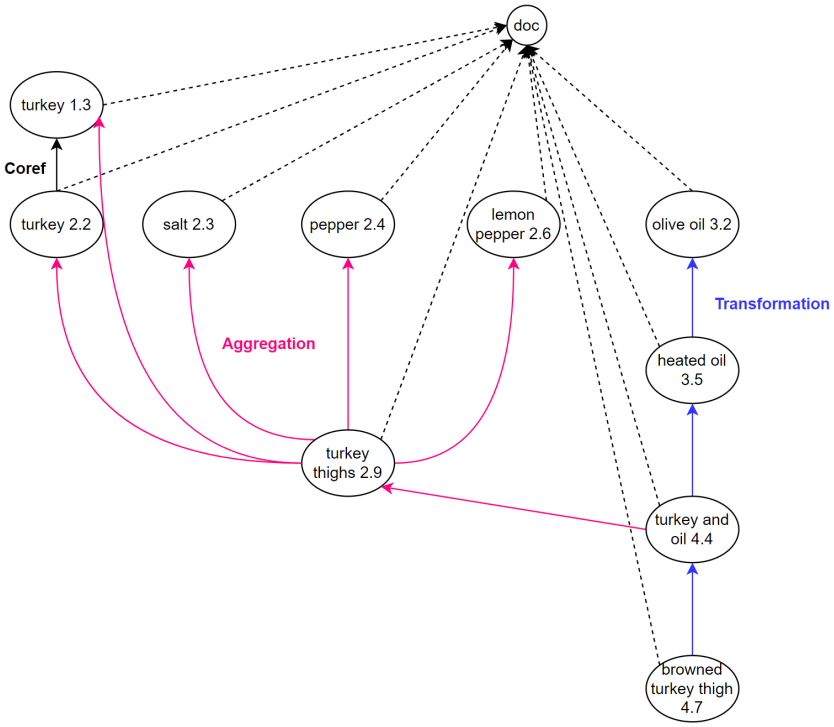

Figure 2: Graph representation of anaphoric relations in a document in CUTL. The black edges are coreference relations. The red edges refer to Aggregation relation while the blue ones refer to Transformation.