On the Impact of Reconstruction and Context for Argument Prediction in Natural Debate

Zlata Kikteva Alexander Trautsch Patrick Katzer Mirko Oest
Steffen Herbold Annette Hautli-Janisz
Faculty of Computer Science and Mathematics
University of Passau
firstname.lastname@uni-passau.de

Abstract

Debate naturalness ranges on a scale from small, highly structured, and topically focused settings to larger, more spontaneous and less constrained environments. The more unconstrained a debate, the more spontaneous speakers act: they build on contextual knowledge and use anaphora or ellipses to construct their arguments. They also use rhetorical devices such as questions and imperatives to support or attack claims. In this paper, we study how the reconstruction of the actual debate contributions, i.e., utterances which contain pronouns, ellipses and fuzzy language, into full-fledged propositions which are interpretable without context impacts the prediction of argument relations and investigate the effect of incorporating contextual information for the task. We work with highly complex spontaneous debates with more than 10 speakers on a wide variety of topics. We find that in contrast to our initial hypothesis, reconstruction does not improve predictions and context only improves them when used in combination with propositions.

1 Introduction

Spontaneous natural debate is anything but easy to track: it contains anaphora, elliptical constructions, fragments, a fuzzy linguistic surface and a wide variety of rhetorical structures. The waters get even murkier when 10+ speakers contribute, multiple, possibly divergent, topics are covered in one debate, the stakes of the interlocutors are high, and debate constraints are low.

So far, debates at this scale of naturalness have been largely ignored in computational argumentation: either the number of participants was restricted and debates were highly constrained (Visser et al., 2020), there was only one topic per debate (Lawrence et al., 2018), or the setting was structured and consisted of monological speaker utterances (Mirkin et al., 2018a).

Most striking is the difference in the underlying data: argument mining on natural debate has either taken propositions as argumentative units of analysis, i.e., fully reconstructed records of speaker contributions that do not need context for interpretation (Gemechu and Reed, 2019; Ruiz-Dolz et al., 2021) or like Lavee et al. (2019) removed claims that contain, for instance, unresolved demonstratives. Another common approach, however, is to take transcripts as is, without any edits or restrictions (Haddadan et al., 2019). Our hypothesis is that using fully reconstructed material, i.e., propositions, increases the performance of argument relation prediction. In the case of locutions, where for instance anaphora and ellipses are not reconstructed, we assume that some of the information relevant to reconstruction is contained within the preceding context, like in an example from the corpus[1] where an anaphoric pronoun *'she'* from the locution '*She's looking at what happened*' can be resolved as *'Sue Grey'* in a proposition '*Sue Gray is looking at what happened*' using preceding context '*Sue Grey is doing this investigation*'.

However, the task of completely reconstructing propositions from locutions, i.e., the actual, skeletal contributions in the debate, is costly: manual reconstruction requires an extensive amount of effort (Hautli-Janisz et al., 2022; Visser et al., 2020), while automatic approach struggles with unresolved non-personal anaphora and omitted verb phrases (Jo et al., 2019, 2020).

Our contributions in this paper are as follows: (1) we provide more insight into model performance in a realistic debate mining setting where only skeletal locutions and not fully reconstructed propositions are available, using the best-performing model on a dataset that is closest in nature to the debates here. Our results indicate that despite the notable structural differences between locutions and propositions, we achieve comparable performance in argument relation prediction for both. (2) We perform a

[1] Node set ID 28238, access via http://ova3.arg.tech

Proceedings of the 10th Workshop on Argument Mining, pages 100–106
December 7, 2023.

detailed error analysis and show that performance across argument relations varies noticeably and that context does not help in solving the issue of using only skeletal locutions but improves the predictions when used in combination with propositions.

2 Related work

In computational argumentation in the debate genre, one strand of research focuses on mostly monological speech, either produced by professional debaters (Mirkin et al., 2018a,b; Lavee et al., 2019; Orbach et al., 2020) or by political actors (Menini et al., 2018). A slightly different variety of debate concerns heavily structured Oxford-style debates (Zhang et al., 2016) where conversational flow is important.

In the case of more natural but still highly moderated debates, the focus is mostly on the political genre with some work on the identification of the central and divisive elements of the debate (Lawrence and Reed, 2017) and prediction of the argument relations using support and attack annotation scheme (Gemechu and Reed, 2019) as well as more complex categories (Ruiz-Dolz et al., 2021). There is more research on the US 2016 elections (Haddadan et al., 2019) as well as the UK Prime-ministerial elections from 2015 (Lippi and Torroni, 2016) with both papers focusing on the detection of argument components such as claims and premises/evidence.

In terms of segmentation, we are similar to most other work in debate mining: Lippi and Torroni (2016) and Haddadan et al. (2019) also assume sentential (or potentially sub-sentential) segments between which argument relations can hold, in contrast to Menini et al. (2018) who seem to take utterances to be the minimal units of analysis. Given the significant amount of argument relations within one utterance, we are confident that the former approach is what captures this genre most appropriately.

3 Empirical basis

3.1 Data

This paper is based on debates in 'Question Time' (QT), a political talk show in the UK broadcasted on BBC1. QT is significantly less structured than debate datasets in previous work, for instance, by Mirkin et al. (2018a,b). In QT, the audience challenges a panel of political figures regarding current topics who then respond and freely discuss the issues with each other. As the participants are different in each episode, their rhetorical skills vary considerably. Topics discussed within and across episodes range from UK-specific and time-sensitive ones such as extension of the lockdowns during the height of the COVID pandemic to more general ones like racism and climate change.

The data is annotated with Inference Anchoring Theory (IAT) (Budzynska et al., 2014, 2016). IAT is a framework that captures how arguments evolve and are reacted to in dialogue, anchoring argument structure in dialogue structure by way of illocutionary connections. The pairs of argumentative units and their relation that are used in this paper have not been annotated in isolation, but have been annotated together with all surrounding material. For the purpose of this paper, we only extract the pairs and their relation (plus the immediately preceding context). Arguments in QT30 comprise inferences ('Inference', supports – serial, divergent, convergent) conflicts ('Conflict', attacks – undercutting, rebutting, undermining) and rephrases ('Rephrase', reformulations of previous content). We extract those argument relations and also include 'No relation' instances (between adjacent units) due to a large number of unconnected contributions in natural debate (see Table 1 for details).

The training data is taken from QT30 (Hautli-Janisz et al., 2022), which comprises analyses of 30 episodes of QT. With 19,842 locutions (plus their propositional counterparts), 280,000 words and more than 10,000 arguments, QT30 is three times larger than the dataset that is most closely related in genre and annotation scheme (Visser et al., 2020). For testing, we use an additional ten episodes of QT on topics that are different than those seen in training (the training data aired between May 2020 and November 2021, test data aired between December 2021 and July 2022). Overall, this leaves us with a training/test split of about 80/20.

Table 1: Number of argument relations of different types and 'No relation' for training and testing

	Training	**Test**	**Total**
Inferences	3,223	845	4,068
Conflicts	697	315	1,012
Rephrases	3,634	1,085	4,719
No relation	4,558	1,052	5,610

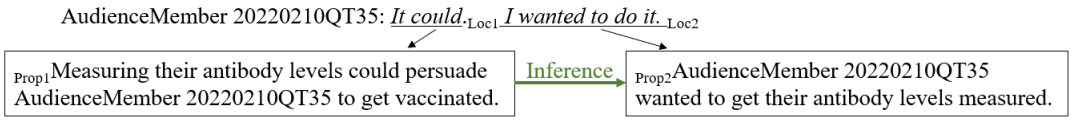


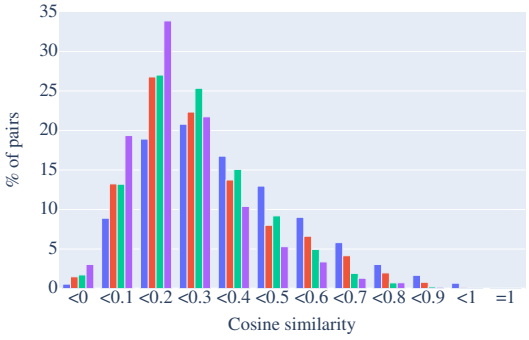
Figure 1: Example with locutions ($Loc1$ and $Loc2$) and propositions ($Prop1$ and $Prop2$)

3.2 Locutions vs. propositions

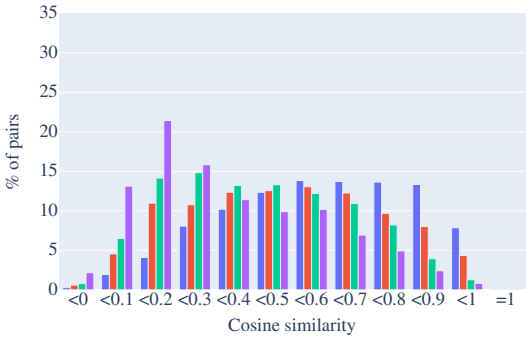
Locutions are the actual, skeletal speaker contributions in a debate. Propositions are their fully reconstructed equivalents: anaphora and ellipses are reconstructed, fragments are transformed into grammatical structures and fuzzy language is resolved. In Figure 1, the locutions of the speaker (an audience member) do not specify what they want to do and why [2]. The manually reconstructed propositions contain this information, namely that the speaker is discussing measuring their antibody levels to inform their decision to get vaccinated. Also, the pronoun *'I'* is reconstructed to *'AudienceMember 20220210QT35'*.

Data extraction We extract the pairs of locutions and matching propositions corresponding to argumentative discourse units (ADUs) which make up an argument (inference, conflict or rephrase) or a 'No relation'. For the 'No relation' category we extract adjacent ADUs which are not connected via an argumentative relation of inference, conflict or rephrase. We also extract the locution and corresponding proposition preceding the first element of the pair – this is what we consider context. This can be an adjacent unit or the one that is dialogically or content-wise preceding the argumentative pair, for instance, in the case of interruptions, the text segment before the interruption is extracted. We end up with a total of 15,409 locution (and the same number of corresponding proposition) triplets.

Structural comparison Locutions and propositions vary consistently in their structure: the average locution length is 11.72 words, propositions tend to be longer with 14.02 words per unit (average Levensthein distance is 19.86, normalized word-level distance is 0.83). For locutions, the average number of pronouns is 1.17 per unit, for propositions it drops significantly to 0.79 per unit. The remaining pronouns in the propositions are either instances where the antecedent of the anaphora is within the same unit (e.g. in *'Boris Johnson lied in his speech about X'*) or cases where their resolution would result in overinterpretation (e.g., *'we need to take care of the older people in care homes'*). Pointing to a similar trend, there are on average 0.37 named entities per locution, compared to 0.76 per proposition.



(a) Locutions



(b) Propositions

Figure 2: Cosine similarity between units in argument relations that are locutions (a) and propositions (b). Rephrases are indicated in blue, conflicts in red, inferences in green, 'No relation' in purple.

Embedding space comparison Given that we use BERT-based prediction for argument relations, we also investigate the impact that the reconstruction has on the embedding space. We calculate the cosine similarity between the first and the second element of the arguments and 'No relation' pairs. We do this for both propositions and locutions us-

[2] Node set ID 27967, access via http://ova3.arg.tech

ing SentenceTransformers with BERT embeddings (the all-MiniLM-L6-v2 model). The results are plotted in Figure 2. The distribution of the cosine similarity in the graphs suggests the following: (1) the model sees very different input in the locution- versus proposition-driven model, as the overall semantic similarity is lower for locutions than propositions while the similarity for propositions is more equally distributed. (2) The semantic space representations can be indicative of the type of relation. Propositions in rephrase relations are more similar than those in conflict or inference ones. The 'No relation' propositions are most dissimilar. Interestingly, units in conflict relations tend to be more similar than inference units.

4 Argument type prediction

4.1 Models

LSTM (baseline) We use softmax activation with categorical cross-entropy as a loss function and the Adam optimizer with a batch size of 32, a maximum sequence length of 200 trained over 4 epochs.

BERT-Based We use pre-trained RoBERTa-large-cased (Liu et al., 2019), the best model identified by Ruiz-Dolz et al. (2021), who worked with the same categories as we do though for more constrained debate settings (fewer topics and speakers, more moderation). In order to compare with a more common BERT model, we also include results for BERT-large-cased. For both models, finetuning is performed on the QT30 data. We use 20% of the training data for validation. For the evaluation, we use 10 extra QT episodes. We train for 6 epochs and choose the best-performing epoch checkpoint on the test data. We use the Adam optimizer with a learning rate of 1e-05, epsilon of 1e-08, a batch size of 32 and a maximum sequence length of 200 which fits our data. In addition, we use 120 warmup steps and a warmup ratio of 0.06. The hyperparameters are taken from Ruiz-Dolz et al. (2021).

4.2 Results

As expected, RoBERTa outperforms the BERT and significantly outperforms the LSTM models [3].

Our best-performing model (Propositions+context) (we use macro F_1-score, as in related work, see Table 2) is still lower in comparison to Ruiz-Dolz et al. (2021), who use the same four-way distinction as we did and achieve the performance of 0.70. However, the corpus they use contains both written discussions as well as transcripts of the US presidential debates which is much more constrained than the debates used here. In comparison to other related work, our results seem to indicate that the less constrained debates are, the lower the performance of the model is. This is supported by the results in earlier work: Menini et al. (2018) who use monological speeches with a binary distinction into inferences and attacks, achieve an F_1-score of 0.82, while Gemechu and Reed (2019) achieve an F_1-score of 0.64 when using a political debate corpus with multiple speakers with the same categories.

Table 2: Macro F_1-scores across models and data

	LSTM	BERT large cased	RoBERTa large cased
Loc	0.25	0.41	0.54
Loc+cont	0.25	0.40	0.53
Prop	0.25	0.41	0.54
Prop+cont	0.24	0.41	**0.56**

Surprisingly, the results indicate that the use of propositions does not improve the performance of the model when compared to the locutions and that context does not help the prediction of relations between locutions, but increases performance when used with propositions. We will reflect on this in the following section. This pattern also holds for the predictions with BERT except for lack of improvement in the case of propositions with context; LSTM also exhibits different kind of behaviour in terms of context but the F_1-scores are too low to be able to draw any meaningful conclusions from them.

5 Error analysis

Context only helps sometimes A closer inspection of the results (for details see Appendix A, Table 3) shows that, with context, the model tends to predict the 'No relation' more often, both in terms of true and false positives. We hypothesize that context locutions in some cases provide information beyond the one relevant for the identification of argument relations thus leading to an increase in the number of predicted 'No relations'.

When we compare the results for propositions with and without context (for details see Appendix A, Table 4), we see that the model is better at pre-

[3] Code available at https://github.com/ZlataKikteva/argmining2023-reconstr

dicting the class of 'Conflict' if context is given, as well as reducing the number of misclassified 'Inference', in particular, inferences misclassified as 'No relation'. With the introduction of context for propositions, we still observe a tendency to overpredict the 'No relation' category, however, this tendency is not as strong as in the case with locutions and context. This can be explained by the fact that due to the reconstruction, the units of the proposition pairs are more likely to have a higher semantic similarity, making it slightly easier for the model to identify the argumentative relations as opposed to 'No relation'.

Reconstruction improves predictions of inferences and rephrases While the F_1-scores of the models based on locutions and propositions are the same, the confusion matrices for the two settings show that the underlying predictions are quite different (the confusion matrices for RoBERTa predictions are attached in Appendix B). The overall tendency when using propositions is leaning towards identifying inferences and rephrases at the cost of 'No relation' (for details see Appendix A, Table 5). Specifically, while the number of correctly predicted 'No relation' propositions went down about 15%, the improvement in the prediction of both rephrases and inferences is about 8%. The example in Figure 1 illustrates the issue: without heavy reconstruction, the model cannot correctly predict the inference and instead goes for 'No relation'. The reconstruction leads to the increased similarity of the embeddings in a number of cases, which makes the prediction of 'Rephrase' and 'Inference' easier while losing out on 'No relation'. In addition to that, this kind of tendency also comes at the cost of misclassifying 'No relation' as inferences.

6 Conclusion

Contrary to our expectations, the reconstruction of skeletal locutions into full-fledged propositions does not necessarily improve the overall performance of the models. What we observe, however, is that the model trained and evaluated on propositions is better at identifying argumentative relations at the cost of the 'No relation' category. In addition, context seems to be beneficial only in the case of propositions as it improves the prediction of conflicts and inferences.

Acknowledgements

The work reported on in this paper was partially funded by the VolkswagenStiftung under grant Az. 98544 'Deliberation Laboratory'.

References

Katarzyna Budzynska, Mathilde Janier, Juyeon Kang, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yaskorska. 2014. Towards argument mining from dialogue. In *Proceedings of the Fifth International Conference on Computational Models of Argument (COMMA 2014)*, pages 185–196. IOS Press.

Katarzyna Budzynska, Mathilde Janier, Chris Reed, and Patrick Saint Dizier. 2016. Theoretical foundations for illocutionary structure parsing. *Argument & Computation*, 7(1):91–108.

Debela Gemechu and Chris Reed. 2019. Decompositional argument mining: A general purpose approach for argument graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 516–526, Florence, Italy. Association for Computational Linguistics.

Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019. Yes, we can! mining arguments in 50 years of US presidential campaign debates. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4684–4690, Florence, Italy. Association for Computational Linguistics.

Annette Hautli-Janisz, Zlata Kikteva, Wassiliki Siskou, Kamila Gorska, Ray Becker, and Chris Reed. 2022. QT30: A corpus of argument and conflict in broadcast debate. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3291–3300, Marseille, France. European Language Resources Association.

Yohan Jo, Jacky Visser, Chris Reed, and Eduard Hovy. 2019. A cascade model for proposition extraction in argumentation. In *Proceedings of the 6th Workshop on Argument Mining*, pages 11–24, Florence, Italy. Association for Computational Linguistics.

Yohan Jo, Jacky Visser, Chris Reed, and Eduard Hovy. 2020. Extracting implicitly asserted propositions in argumentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 24–38, Online. Association for Computational Linguistics.

Tamar Lavee, Matan Orbach, Lili Kotlerman, Yoav Kantor, Shai Gretz, Lena Dankin, Michal Jacovi, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. 2019. Towards effective rebuttal: Listening comprehension using corpus-wide claim mining. In *Proceedings of the 6th Workshop on Argument Mining*, pages 58–66, Florence, Italy. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2017. Using complex argumentative interactions to reconstruct the argumentative structure of large-scale debates. In *Proceedings of the 4th Workshop on Argument Mining*, pages 108–117, Copenhagen, Denmark. Association for Computational Linguistics.

John Lawrence, Jacky Visser, and Chris Reed. 2018. Bbc moral maze: Test your argument. In *7th International Conference on Computational Models of Argument, COMMA 2018*, pages 465–466. IOS Press.

Marco Lippi and Paolo Torroni. 2016. Argument mining from speech: Detecting claims in political debates. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

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.

Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2018. Never retreat, never retract: Argumentation analysis for political speeches. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Shachar Mirkin, Michal Jacovi, Tamar Lavee, Hong-Kwang Kuo, Samuel Thomas, Leslie Sager, Lili Kotlerman, Elad Venezian, and Noam Slonim. 2018a. A recorded debating dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Shachar Mirkin, Guy Moshkowich, Matan Orbach, Lili Kotlerman, Yoav Kantor, Tamar Lavee, Michal Jacovi, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. 2018b. Listening comprehension over argumentative content. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 719–724, Brussels, Belgium. Association for Computational Linguistics.

Matan Orbach, Yonatan Bilu, Assaf Toledo, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2020. Out of the echo chamber: Detecting countering debate speeches. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7073–7086, Online. Association for Computational Linguistics.

Ramon Ruiz-Dolz, Jose Alemany, Stella M. Heras Barberá, and Ana García-Fornes. 2021. Transformer-based models for automatic identification of argument relations: A cross-domain evaluation. *IEEE Intelligent Systems*, 36(6):62–70.

Jacky Visser, Barbara Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzynska, and Chris Reed. 2020. Argumentation in the 2016 US presidential elections: annotated corpora of television debates and social media reaction. *Language Resources and Evaluation*, 54(1):123–154.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational flow in Oxford-style debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 136–141, San Diego, California. Association for Computational Linguistics.

A Differences in class assignments (RoBERTa-large-cased predictions)

We generate these tables based on the class predictions from the confusion matrices.

Table 3: Difference in class assignments between locutions and locutions with context (in percentage)

	Inference	Conflict	Rephrase	No rel.
Inference	-6.75%	0.36%	-0.59%	6.98%
Conflict	-8.57%	-1.59%	-0.63%	10.79%
Rephrase	-1.94%	0.09%	-2.58%	4.42%
No rel.	-5.32%	-0.48%	-1.24%	7.03%

Table 4: Difference in class assignments between propositions and propositions with context (in percentage)

	Inference	Conflict	Rephrase	No rel.
Inference	-1.30%	-0.83%	-0.12%	2.25%
Conflict	-6.67%	1.90%	-2.54%	7.30%
Rephrase	-2.86%	-0.28%	-0.65%	3.78%
No rel.	-10.17%	-0.57%	3.14%	7.60%

Table 5: Difference in class assignments between locutions and propositions (in percentage)

	Inference	Conflict	Rephrase	No rel.
Inference	7.69%	-1.18%	2.96%	-9.47%
Conflict	7.30%	1.27%	2.86%	-11.43%
Rephrase	-0.09%	0.28%	8.48%	-8.66%
No rel.	14.16%	0.19%	0.86%	-15.21%

B Confusion matrices (RoBERTa-large-cased predictions)