

Breakdown Detection in Negotiation Dialogues (Student Abstract)

Atsuki Yamaguchi,^{1,2} Katsuhide Fujita²

¹The University of Sheffield

²Tokyo University of Agriculture and Technology*
ayamaguchi1@sheffield.ac.uk, katfuji@cc.tuat.ac.jp

Abstract

In human-human negotiation, reaching a rational agreement can be difficult, and unfortunately, the negotiations sometimes break down because of conflicts of interests. If artificial intelligence can play a role in assisting with human-human negotiation, it can assist in avoiding negotiation breakdown, leading to a rational agreement. Therefore, this study focuses on end-to-end tasks for predicting the outcome of a negotiation dialogue in natural language. Our task is modeled using a gated recurrent unit and a pre-trained language model: BERT as the baseline. Experimental results demonstrate that the proposed tasks are feasible on two negotiation dialogue datasets, and that signs of a breakdown can be detected in the early stages using the baselines even if the models are used in a partial dialogue history.

Introduction

Bargaining happens daily in society. Although negotiators generally aim to maximize their profits, exceptions exist where they cannot consider every potential agreement because of surplus options; moreover, their feelings and social background can adversely affect the benefits of them all.

Researchers have previously attempted handling human negotiations using automated negotiating agents to achieve a rational agreement (CatholijnJonker et al. 2017). However, these agents can rarely resolve real-world negotiation problems because of their complexity, and moreover, it is difficult to model the preferences and strategies of human negotiators manually, which is a requirement in automated negotiation. Therefore, there is a need to model these negotiations in an end-to-end manner without manual intervention.

With the emergence of neural network-based natural language processing, Lewis et al. (2017) and He et al. (2018) have proposed end-to-end neural negotiating models that can bargain on goods in natural language. While previous efforts aimed to develop an agent that can negotiate in natural language, few studies have been conducted on supporting human-human negotiation. Only Iwasa and Fujita (2018) introduced an end-to-end model that can estimate the preferences of each agent trained on the dataset reported by Lewis

et al. (2017) and tried to find the Nash bargaining solution for each dialogue.

Hence, this paper focuses on new tasks for predicting breakdowns based on negotiation dialogues in an end-to-end manner for supporting human-human negotiation. High performance in these tasks enables a model to detect signs of a breakdown beforehand and to provide feedback for the negotiators, which contributes toward the prevention of shortage in negotiations and the loss of social welfare.

Tasks and Problem Settings

Let D be a negotiation dialogue comprising $n \in \mathbb{N}$ dialogue turns $\{s_1, s_2, \dots, s_n\}$, where s is a message from one of the negotiators and has one or more sentences. In the task of breakdown detection in negotiation dialogues, D serves as an input, while its outcome $t = \{0, 1\}$ is an output¹. We regard $t = 0$ as a successful dialogue, while $t = 1$ indicates a dialogue breakdown.

We also propose an early breakdown detection task, where the number of turns in D is limited to a specified ratio and subsequently fed into a classification model to predict the outcome. To limit the number of turns in a dialogue, we calculate the number of turns n in D and extract D based on a predefined ratio $r \in \{0.25, 0.5, 0.75, 1.0\}$. The number of extracted turns $n' \in \mathbb{N}$ is determined by rounding $n \times r$.

Experiments

Datasets We considered two negotiation dialogue datasets, DEALORNODEAL (DN) (Lewis et al. 2017) and CRAIGSLISTBARGAIN (CB) (He et al. 2018), for the proposed tasks. The percentage of agreement in the DN and CB datasets was 76.2 and 74.9, respectively.

Baseline Models We tested four baseline models to verify the feasibility of the proposed tasks: GRU+SA (Iwasa and Fujita 2018), GRU+SA w/ ELMo, BERT_{BASE} (Devlin et al. 2019)², and DistilBERT_{BASE}².

¹Negotiation datasets released by Lewis et al.; He et al. (2017; 2018) have an outcome label (agree or disagree) for each dialogue.

²We utilized the BERT implementations of HuggingFace. <https://github.com/huggingface/transformers>

*2-14-16, Nakacho, Koganei, Tokyo, Japan, +81423887141
Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Model	CRAIGSLISTBARGAIN					DEALORNODEAL				
	AUC	TN	TP	FP	FN	AUC	TN	TP	FP	FN
GRU+SA	0.85±0.02	0.93±0.03	0.77±0.06	0.07	0.23	0.77±0.02	0.97±0.01	0.58±0.04	0.03	0.42
GRU+SA w/ ELMo	0.86±0.03	0.94±0.01	0.78±0.07	0.06	0.22	0.77±0.01	0.98±0.01	0.56±0.03	0.02	0.44
BERT _{BASE}	0.85±0.02	0.94±0.02	0.77±0.04	0.06	0.23	0.79±0.02	0.97±0.01	0.62±0.03	0.03	0.38
DistilBERT _{BASE}	0.84±0.03	0.92±0.02	0.77±0.07	0.08	0.23	0.77±0.02	0.98±0.01	0.56±0.04	0.02	0.44

Table 1: Results of breakdown detection in CB and DN datasets. “SA” denotes self-attention mechanism. The values of the confusion matrix were normalized on a set each of TN and FP and TP and FN.

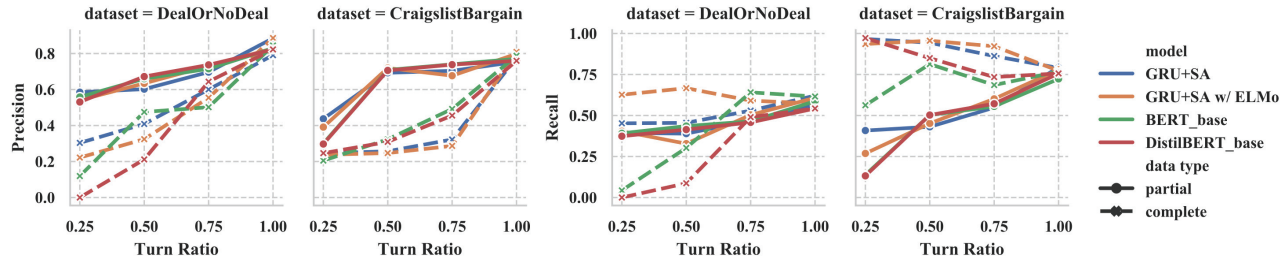


Figure 1: Results of early breakdown detection in CB and DN datasets. The horizontal axis denotes the ratio of turns to be preserved. The dotted line labeled “complete” indicates the results obtained when the number of turns in the validation data is limited, while the solid line labeled “partial” indicates the results obtained when the number of turns in both training and validation data is limited.

Training Details We applied stratified five-fold cross-validation when running each baseline model. Early stopping was adopted with the patience value set to three. The models were trained to minimize binary cross-entropy loss.

Results The ROC-AUC values in both datasets demonstrate that the task of breakdown detection is feasible because these values are significantly higher than those of random classification. Although we achieved high accuracy in predicting a successful dialogue, classification of a failure case remains a challenge (Table 1: TP and TN).

Figure 1 shows the results of early breakdown detection in the baseline models. We can observe that the precision of models trained on limited data is higher compared with the models trained on non-limited data, while the recall values exhibit the opposite trend, with the exception of the BERT-based models in the DN dataset whose training may be unstable due to insufficient data. The potential causes of these differences are as follows: (1) By limiting the number of turns in the training data, the model may capture specific features related to negotiation breakdown more easily in a shorter dialogue. (2) The models trained on non-limited data regard the length of a dialogue as an important feature, and as a result, classify short dialogues as breakdown cases.

Notably, when we feed more than half of the turns in a negotiation dialogue into the model trained on limited data, a breakdown dialogue can be more accurately classified, with more than 70% precision in both datasets. In the task of breakdown detection, it is important to achieve high precision even if the recall is low because the model should avoid classifying a successful dialogue as a breakdown case. Therefore, these results suggest that it is better to limit the number of turns in the training data to detect early signs of breakdown. In the future, the development of a method that

improves recall while maintaining precision is an important research direction for the task of early breakdown detection.

Conclusion and Future Work

We have introduced new tasks that predict a negotiation breakdown using an end-to-end approach for supporting human-human negotiation. We found that breakdown detection is feasible, and the number of turns of training data should be limited for early breakdown detection. In future work, we aim to develop a new evaluation metric to consider the quality of each negotiation dialogue.

Acknowledgments

This work was supported by JST CREST (JPMJCR15E1).

References

- CatholijnJonker; ReyhanAydogan; TimBaarslag; KatsuhideFujita; TakayukiIto; and KoenHindriks. 2017. Automated negotiating agents competition (anac). In *Proceedings of AAAI 2017*, 5070–5072.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019, Volume 1*, 4171–4186.
- He, H.; Chen, D.; Balakrishnan, A.; and Liang, P. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of EMNLP 2018*, 2333–2343.
- Iwasa, K., and Fujita, K. 2018. Prediction of nash bargaining solution in negotiation dialogue. In *Proceedings of PRICAI 2018*, 786–796.
- Lewis, M.; Yarats, D.; Dauphin, Y.; Parikh, D.; and Batra, D. 2017. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of EMNLP 2017*, 2443–2453.