Improving Abstractive Dialogue Summarization with Speaker-Aware Supervised Contrastive Learning

Anonymous ACL submission

Abstract

Pre-trained models have brought remarkable success on the text summarization task. For dialogue summarization, the subdomain of text summarization, utterances are concatenated to flat text before being processed. As a result, existing summarization systems based on pre-trained models are unable to recognize the unique format of the speaker-utterance pair well in the dialogue. To investigate this issue, we conduct probing tests and manual analysis, and find that the powerful pre-trained model can not identify different speakers well in the conversation, which leads to various factual errors. Moreover, we propose three speaker-aware supervised contrastive learning (SCL) tasks: *Token-level SCL*, *Turn-level SCL*, and *Global-level SCL*. Comprehensive experiments demonstrate that our methods achieve significant performance improvement on two mainstream dialogue summarization datasets. According to detailed human evaluations, pre-trained models equipped with SCL tasks effectively generate summaries with better factual consistency.

1 Introduction

Dialogue summarization aims to condense the essential information in the dialogue into a brief text. Compared with text summarization, the conversations are semi-structured data and contain multiple participants who shall be distinguished (Gurevych and Strube, 2004; Feng et al., 2021a). Furthermore, dialogues combine features such as informal language, coreference, and repetition (Chen and Yang, 2020). All of these bring new challenges to the existing text summarization methods.

Although pre-trained models have achieved great success in text summarization (Liu and Lapata, 2019; Lewis et al., 2020; Zhang et al., 2020), how to properly utilize them in dialogues with a special speaker-utterance structure is still an obstacle. A line of previous work utilizes pre-trained models and deals with dialogue summarization as flat text. Chen and Yang (2020) segment dialogues into blocks from multiple semantic views and process them using BART. Feng et al. (2021b) use DialoGPT (Zhang et al., 2019) as an unsupervised annotator to help models understanding conversations. However, due to the gap with the pre-training object, the pre-trained models are hard to capture speaker information. To investigate these, we conduct a manual analysis on a popular dataset SAMSum (Gliwa et al., 2019) and discover that, even for the state-of-the-art model BART, **50%** of the generated summaries contain factual errors for dialogues with multiple speakers. Among them, up to **68.4%** are caused directly by speaker confusion and speaker missing (see Section 3.2). As shown in Table 1, the model's inability to identify speakers results in serious factual inconsistencies.

Dialogue Text
Jeff: Should we go to the village party? Lia: I'm too tired after hiking. Mico: I'd like to go, there may be some hot boys! Lia: I doubt Jim: like a real village boy? Jim: who doesn't even speak English? Mico: yes, the dummer, the better. Jim: haha, stupid fucks good, they say. Mico: I confirm! Lia: not my cup of tea. Mico: I'll go there, who wants to join? Jeff: I'll go as well. Mico: wanna drive? Jeff: so you could drink? Mico: would be nice, hahah. Jeff: not excited, but ok.
Gold Summary
Mico and Jeff will go to the village party. Jeff will drive.
Baseline Summary (by BART)
Jeff, Lia and Mico are going to the village party. Lia is too tired to go. Mico will drive.
Our Summary
Jeff and Mico are going to the village party. Lia is too tired after hiking. Jeff will drive.

Table 1: A dialogue example in the SAMSum dataset. The summary generated by BART has two factual errors: Lia is not going to the village party; it will be Jeff driving instead of Mico. Our model can generate factually correct summaries.

Another tributary of previous work (Zhao et al., 2019; Liu and Chen, 2019; Zhu et al., 2020; Lei et al., 2021) utilizes the hierarchical network instead of pre-trained models to leverage the dialogue's structural information. However, how to explicitly model the information of speakers in pre-trained sequence-to-sequence (seq2seq) models remains unsolved. Zhu et al. (2020) introduces speaker embedding to distinguish speakers for meetings with fixed participants. However, in most cases, the number and identity of the participants in the conversations are unknown. Thus the trained embedding is not a general solution.

Intuitively, if the representation derived from the encoder has sufficient information to identify speakers, the decoder will produce superior summaries, especially for summaries that follow a pattern of *someone does something* as shown in Table 1. In this paper, we first conduct a probing experiment to show that the representation of the dialogue obtained from BART can not distinguish speakers well. To address this issue, we use contrastive learning to improve the alignment of the representation derived from the encoder, i.e., to make the encoder output diverse hidden states based on corresponding speakers. We propose three speaker-aware supervised contrastive learning tasks: *Token-level SCL*, *Turn-level SCL*, and *Global-level SCL*. By jointly training these tasks in the fine-tuning stage, we can substantially improve the model's ability to identify different speakers and further understand the content of the whole dialogue. Comprehensive experiments and human evaluations on SAMSum and AMI (McCowan et al., 2005) reveal that our models generate summaries with higher ROUGE scores and better factual consistency. Our main contributions include (a) this is the first work to give a detailed investigation of the speaker identification problem in dialogue summarization, (b) proposing speaker-aware SCL tasks to address the problem, and evaluating our methods with the experimental and manual examination.

2 Method

2.1 Probing Test

To investigate how well pre-trained seq2seq models can distinguish speakers, we conduct a simple probing experiment on SAMSum, a widely-used dialogue summarization corpus. Concretely, we first encode the dialogue text with the BART (Lewis et al., 2020) encoder and randomly sample K tokens to obtain their hidden states. Then, in pairs, we aggregate and feed these hidden states into MLP to determine whether they are from the same speaker. We train the MLP layers on the SAMSum training set (in other words, we freeze the parameters of the BART encoder during the training stage) and then assess the classification accuracy on the test set. For vanilla BART, the accuracy is **58.1%**. After fine-tuning BART with the summarization task on SAMSum before the probing test (parameters are not frozen), the accuracy is still only **60.2%**[1]. Given the task's simplicity, this result indicates that pre-trained seq2seq models can not identify speakers well from flat dialogue text.

2.2 Supervised Contrastive Learning Tasks

To address the above problem, inspired by the research about contrastive learning (Mikolov et al., 2013; Saunshi et al., 2019; He et al., 2020; Velickovic et al., 2019), we introduce SCL tasks during the fine-tuning stage to minimize the distance between representations of utterances from the same speaker and vice versa.

Formally, a dialogue $D = (t_1, t_2, \cdots, t_n)$ consists of n turns, and each turn t_i contains the utterance u_i and the corresponding speaker s_i, that is, $t_i = (s_i, u_i)$. Firstly, we use Transformer-Encoder (Vaswani et al., 2017) to model the dialogue-level contextual representation of each tokens.

$$H = \text{Transformer-Encoder}(D), \tag{1}$$

where the input sequence is the concatenation of all turns. Then, we can generate the summary $\hat{y}$ with Transformer-Decoder. The generation loss $\mathcal{L}_{gen}$ is cross-entropy loss between $\hat{y}$ and gold summary y.

Incorporating Contrastive Loss To enable the utterance representation to contain more speaker information, we incorporate three levels of contrastive losses.

Generally, let (o_i, s_i) denote a sampled token or a sampled utterance and its associative speaker. The contrastive loss $\mathcal{L}_{ctr}$ for the SCL task is calculated as follows:

$$\mathcal{L}_+ = \sum_{i,j}^{s_i=s_j} -\log(\sigma(\mathbf{o}_i \cdot \mathbf{o}_j)), \tag{2}$$

$$\mathcal{L}_- = \sum_{i,j}^{s_i \neq s_j} -\log(1 - \sigma(\mathbf{o}_i \cdot \mathbf{o}_j)), \tag{3}$$

$$\mathcal{L}_{ctr} = \mathcal{L}_+ + \mathcal{L}_-, \tag{4}$$

where σ is logistic function that measures the similarity between two representations and $\mathbf{o}_i$ is the

[1] By jointly training the Global-level SCL task in fine-tuning stage, the accuracy reaches **77.9%**.

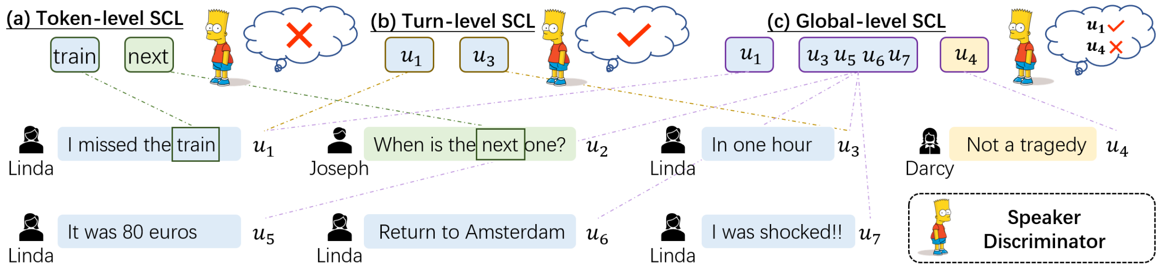

Figure 1: Overview of our speak-aware SCL tasks. *Token-level SCL* and *Turn-level SCL* mean the model needs to discriminate whether two tokens/turns are from the same speaker. *Global-level SCL* let the model choose what the speaker might say in a particular turn when given all the utterances of this speaker. The representations are obtained by inputting the whole dialogue into the encoder.

contextual representation of o_i derived from H. The detailed definitions of o_i are in Section 2.2.1 $\sim$ 2.2.3.

The final loss $\mathcal{L} = \lambda\mathcal{L}_{ctr} + \mathcal{L}_{gen}$ and λ is the weight coefficient to adjust the ratio of $\mathcal{L}_{ctr}$ and $\mathcal{L}_{gen}$ in the final loss $\mathcal{L}$. The model is supposed to maximize similarity among samples of the same speaker and vice versa while optimizing for the summary generation.

2.2.1 Token-level SCL

The first task is the Token-level SCL which means the model distinguishes whether two tokens are from the same speaker. As illustrated in Figure 1(a), we randomly sample m token-speaker pairs $T = \{(o_1, s_1), (o_2, s_2), ..., (o_m, s_m)\}$ from D, where o_i is a token and s_i is the corresponding speaker. The hidden state of o_i obtained through the encoder is used to represent the i-th sample.

2.2.2 Turn-level SCL

Compared with Token-level SCL, we increase the granularity of the input to fuse the semantic information of the context. As shown in Figure 1(b), we randomly sample two turns from D and mask the speaker names in text, denoted as (o_i, s_i) and (o_j, s_j). Then we derive $\mathbf{o}_i$ by taking the mean pooling of the hidden states of all tokens in o_i.

2.2.3 Global-level SCL

To maximize the mutual information between utterances of the same speaker (Linsker, 1988; Kong et al., 2019), we extend the Turn-level SCL task to Global-level SCL by introducing global information. Intuitively, we can understand the speaking style of a specific person from all the words he or she has said. Therefore, we provide the model with all the utterances of a certain speaker and let it choose what this speaker might say in a particular turn (described in Figure 1(c)). Concretely, we first mask all the speaker names and randomly sample a speaker whose utterances set $\tilde{S}_i$ which has at least two elements. Among $\tilde{S}_i$, we randomly choose a utterance (o_i, s_i) as the positive sample, and randomly choose another utterance (o_j, s_j) from $D - \tilde{S}_i$ as the negative sample. Thus the global utterance sample of this speaker is $(\tilde{S}_i - o_i, s_i)$. The model is supposed to maximize the mutual information between the representation of the global sample and the positive sample, and vice versa. The representations are derived from mean pooling, the same as what we do in Turn-level SCL. In contrast to Turn-level SCL, Global-level SCL needs the model's overall comprehension of the dialogue-format context.

3 Experiment

In this section, we conduct experiments and human evaluations on the popular datasets SAMSum and AMI. More descriptions of the datasets and the implementation details can be found in the Appendix.

3.1 Experimental Result and Analysis

We provide several latest strong seq2seq models as baselines, including PGNet (See et al., 2017), UniLM (Dong et al., 2019) and BART+DialoGPT (Feng et al., 2021b) in the first part of Table 2. Following previous settings (Gliwa et al., 2019; Feng et al., 2021a), we use *py-rouge*[2] package for evaluation on SAMsum and use *pyrouge*[3] on AMI. Experimentally, our models obtain clear improvement on both two datasets compared to the BART baseline, and achieve the state-of-the-art result on SAMSum.

[2] https://pypi.org/project/py-rouge/
[3] https://github.com/bheinzerling/pyrouge

Model	SAMSum			AMI		
	R-1	R-2	R-L	R-1	R-2	R-L
PGNet (See et al., 2017)	40.08	15.28	36.63	42.60	14.01	22.62
UniLM (Dong et al., 2019; Zhu et al., 2021)	50.00	26.03	42.34	50.61	**19.33**	25.06
Multi-view BART (Chen and Yang, 2020)	53.42	27.98	49.97	-	-	-
BART+DialoGPT (Feng et al., 2021b)	53.70	28.79	50.81	-	-	-
PGN+DialoGPT (Feng et al., 2021b)	-	-	-	50.91	17.75	24.59
BART	53.01	28.05	49.89	50.67	17.18	24.96
BART + Token-level SCL task	53.85	29.21	50.94	51.03	17.23	25.21
BART + Turn-level SCL task	54.12	29.53	51.10	51.15	17.85	**25.45**
BART + Global-level SCL task	**54.22**	**29.87**	**51.35**	**51.40**	17.81	25.30

Table 2: Results on the test sets of SAMSum and AMI, and "R" is short for "ROUGE".

Specific to the three tasks, the improvement brought by *Token-level SCL* is relatively tiny. The reason may be that the utilization of positional information is enough for BART to optimize the contrastive loss for two tokens. For *Turn-level SCL* and *Global-level SCL*, the pooling layer reduces the impact of position embedding, thereby forcing the model to focus on the semantic information of the utterances. Therefore, the model can further capture the characteristics of the dialogue data. *Global-level SCL* performs best in both datasets, which illustrates that when the model has global perspectives for each speaker, it can enhance the model's comprehension of the whole dialogue.

3.2 Human Evaluation

We also conduct human evaluations to investigate if our method leads to fewer factual errors. Automatic metrics like FACTCC (Kryściński et al., 2019) are not used since the neural-model-based metrics perform poorly in dialogue data due to the significant domain gap. And most of the factual errors in the dialogue summarization are caused by misidentification of the speaker, which can not be reflected by automatic metrics. Here we use BART and BART with the Global-level SCL task for comparison.

Error Types Firstly, we divide the factual errors into three categories manually:(a) **Speaker Confusion**: Model confuses speakers participating in a specific event; (b) **Speaker Missing**: A speaker is mentioned in the gold summary, while the model hits the event but misses this speaker. (c) **Semantic Error**: Errors caused by a misunderstanding of semantics, and they are not directly related to any speakers. More cases about the error types can be found in Appendix.

Model	BART	BART + Global SCL
Speaker Confusion	9	5
Speaker Missing	30	20
Semantic Errors	18	16

Table 3: The number of factual errors for the baseline model and our model on the SAMSum dataset.

Result We evaluate 100 dialogues from the test set of SAMsum. For SAMSum, we choose dialogues with at least three speakers to explore the model's ability to understand multi-person interaction, and the result is shown in Table 3. **50%** of the summaries generated by BART contain factual errors, of which **68.4%** are related to the Speaker Confusion or Speaker Missing[4]. In comparison, our model decreases the number of speaker-related factual errors by **35.9%**. The SCL task helps the model to better distinguish the speakers and intuitively reduce the confusion. Furthermore, in many cases, speakers whose utterances account for a small proportion will be missed by baseline model. With our method, the representation of their utterances are different from others, reducing the missing error. The detailed human evaluation on AMI is listed in Appendix.

4 Conclusion

In this paper, we focus on the speaker identification problem in the dialogue summarization task. Through the probing test and manual analysis, we find that the existing pre-trained model can not identify different speakers well in the conversation, leading to factual errors. Therefore, we propose three speaker-aware SCL tasks to address this problem. Experimental results and human evaluations illustrate the effectiveness of our methods.

[4]Please note that there may be multiple types of errors in a single sample.

5 Ethical Considerations

For human evaluation in section 3.2, we recruited two annotators to see if there are any factual inconsistencies in generated summaries. The generators of all summaries are hidden from the annotators to avoid any subjective bias in our proposed methods. For the SAMsum dataset, we give more priority to dialogues with more speakers and adopt a random strategy when the numbers of speakers are same. Furthermore, both annotators were compensated fairly.

References

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118, Online. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *arXiv preprint arXiv:1905.03197*.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021a. A survey on dialogue summarization: Recent advances and new frontiers. *arXiv preprint arXiv:2107.03175*.

Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. 2021b. Language model as an annotator: Exploring dialogpt for dialogue summarization. *arXiv preprint arXiv:2105.12544*.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Iryna Gurevych and Michael Strube. 2004. Semantic similarity applied to spoken dialogue summarization. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 764–770.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

Lingpeng Kong, Cyprien de Masson d'Autume, Wang Ling, Lei Yu, Zihang Dai, and Dani Yogatama. 2019. A mutual information maximization perspective of language representation learning. *arXiv preprint arXiv:1910.08350*.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

Yuejie Lei, Yuanmeng Yan, Zhiyuan Zeng, Keqing He, Ximing Zhang, and Weiran Xu. 2021. Hierarchical speaker-aware sequence-to-sequence model for dialogue summarization. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7823–7827. IEEE.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612.

Ralph Linsker. 1988. Self-organization in a perceptual network. *Computer*, 21(3):105–117.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 3721–3731.

Zhengyuan Liu and Nancy Chen. 2019. Reading turn by turn: Hierarchical attention architecture for spoken dialogue comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5460–5466.

Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, M Guillemot, Thomas Hain, J Kadlec, Vasilis Karaiskos, et al. 2005. The ami meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, volume 88, page 100. Citeseer.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. 2019. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, pages 5628–5637. PMLR.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Petar Velickovic, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. 2019. Deep graph infomax. In *ICLR (Poster)*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Zhou Zhao, Haojie Pan, Changjie Fan, Yan Liu, Linlin Li, Min Yang, and Deng Cai. 2019. Abstractive meeting summarization via hierarchical adaptive segmental network learning. In *The World Wide Web Conference*, pages 3455–3461.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. Mediasum: A large-scale media interview dataset for dialogue summarization. *arXiv preprint arXiv:2103.06410*.

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 194–203, Online. Association for Computational Linguistics.

A Datasets

We apply our methods on the large version of BART and evaluate our model on SAMSum and AMI datasets using ROUGE score (Lin and Och, 2004). SAMSum consists of 16,369 samples with an average of 2.4 participants and 83.9 words. AMI consists of 137 meeting records of four fixed speakers, which have 4,757 words on average. Due to the limitation of our computing resources, all our inputs are truncated to 1,024 tokens. We use the same split as Gliwa et al. (2019) and Zhu et al. (2020) for SAMSum and AMI, respectively.

B Implementation Details

Hyperparameters	**SAMSum**	**AMI**
Batch Size	8	24
Total Steps	10,000	600
Eval Steps	1,000	20
Learning Rate	[2e-5,3e-5]	[2e-5,3e-5]
Label Smoothing Factor	0.1	0
Warm-up Type	linear	linear
Warm-up Steps	0	100
Max Target Length	128	300

Table 4: Hyperparameters we used for fine-tuning BART on SAMSum and AMI.

Some of our hyperparameters are listed in Table 4. Other hyperparameters are the same as the default of *facebook/bart-large* of transformers[5]. The weight coefficient factor λ is searched from $\{0.01, 0.001\}$. It takes up to 2 hours for one run on SAMSum or AMI using one GeForce RTX 3090.

We use the validation set to select the best checkpoint, and evaluate the checkpoint on the test set.

C Human Evaluation on AMI

Model	BART	BART + Global SCL
Speaker Confusion	3	3
Speaker Missing	3	0
Semantic Errors	18	14

Table 5: The number of factual errors for the baseline model and our model on the AMI dataset.

For the AMI dataset, we evaluate all 20 samples of the test set, and the result is shown in Figure 5. All speakers in AMI are fixed, so the two models rarely confuse them. However, as a result of the truncation, some utterances are only left a small part and become easily overlooked. With our method, BART can better identify the corresponding speakers. Due to the loss of input information, both models have a large number of semantic errors. Compared to the baseline model, our model decreases the number by **22.2%**.

[5] https://github.com/huggingface/transformers

D Case Study

In order to better illustrate the three types of errors mentioned in Section 3.2, we provide more cases here. An example of confusing speakers is shown in Table 1 of the main paper. Examples of missing speakers and semantic errors are shown in Table 6.

Dialogue Text 1
Ann: Congratulations!! Ann: You did great, both of you! Sue: Thanks, Ann Julie: I'm glad it's over! Julie: That's co cute of you, girl! Ann: Let's have a little celebration tonight! Sue: I'm in Julie: me too!!! aww
Gold Summary 1
Ann, Sue and Julie did a great job and they will have a little celebration tonight.
Baseline Summary 1 by BART
Sue and Julie are going to celebrate their success tonight.
Our Summary 1
Ann, Sue and Julie are celebrating their wins.
Dialogue Text 2
Sarah: omg Laura! sorry you didn't get any replies!!! Did you manage? Laura: hahaha! Awksssss... no worries, I solved it Sarah: awkward silence <crickets> Laura: hahaha no it's all good really!! Raf: Laura, I'm so sorry!!! been so swamped, totally forgot to text you back! where are you?? Sarah: Exotic little island called Linate :D Laura: Sarah which hotel are you at??? I'm here too!!!
Gold Summary 2
Neither Raf nor Sarah remembered to reply to Laura but she managed anyway. Both Sarah and Laura are in Linate.
Baseline Summary 2 by BART
Laura didn't get any replies to Sarah's messages.Laura is on an island called Linate. Laura and Sarah are staying at the same hotel.
Our Summary 2
Laura didn't get any replies from Sarah and Raf. Sarah and Laura are on an exotic little island called Linate.

Table 6: Sample 1 is an example about the speaker missing error. The summary generated by BART misses Ann. The dialogue sample is from the SAMSum dataset. Sample 2 is an example about the semantic error and the speaker missing error. The summary generated by BART misses Raf (Speaker Missing Error), and makes it out of thin air that Sarah and Laura are staying at the same hotel (Semantic Error). All samples are from the SAMsum dataset.