Capturing Speaker Incorrectness: Speaker-Focused Post-Correction for Abstractive Dialogue Summarization

Dongyub Lee[1] **Jungwoo Lim**[2*] **Taesun Whang**[3*] **Chanhee Lee**[2]
Seungwoo Cho[4] **Mingun Pak**[5] **Heuiseok Lim**[2]
[1]Kakao Corp.
[2]Korea University
[3]Wisenut Inc.
[4]Kakao Enterprise
[5]Microsoft

Abstract

In this paper, we focus on improving the quality of the summary generated by neural abstractive dialogue summarization systems. Even though pre-trained language models generate well-constructed and promising results, it is still challenging to summarize the conversation of multiple participants since the summary should include a description of the overall situation and the actions of each speaker. This paper proposes self-supervised strategies for speaker-focused post-correction in abstractive dialogue summarization. Specifically, our model first discriminates which type of speaker correction is required in a draft summary and then generates a revised summary according to the required type. Experimental results show that our proposed method adequately corrects the draft summaries, and the revised summaries are significantly improved in both quantitative and qualitative evaluations.

1 Introduction

Research on abstractive dialogue summarization recently have achieved remarkable improvements (Chen and Yang, 2020; Malykh et al., 2020; Chen and Yang, 2021; Wu et al., 2021; Zhu et al., 2021) in diverse domains such as daily dialogue, interview, and movie. Despite the promising performance of the pre-trained language models (e.g., UniLM (Dong et al., 2019) and BART (Lewis et al., 2020)), the capability of summarizing the multi-party conversation is still limited. Due to its difficulty of considering all the actions of every speaker for describing a scene is quite challenging. Specifically, Chen and Yang (2020) emphasize dealing with multi-party situation in dialogue summarization based on several criteria, such as role & language change, referral & coreference, and multiple turns & participants.

Dialogue 1 (D1)
Isabella: Hi Betty! Isabella: It was very nice to listen about your work yesterday. Thank you for sharing that! Isabella: If you wanted to do sth together, let me know. Betty: Thank you! **Draft Summary (BART$_{base}$)** ***Betty*** *was listening to* ***Isabella****'s work yesterday. If she wanted to do something together, she should let her know.*
Dialogue 2 (D2)
Molly: listen I've got a free ticket to the Muse concert in Cracow, want to come with me? Hannah: nah, I don't like them Molly: what about you Anna Anna: yassss please. let's go! <3" **Draft Summary (BART$_{base}$)** *Molly has a free ticket to the Muse concert in Cracow. Hannah and* ***Anna*** *don't like them.*

Table 1: Examples of the incorrect summaries that contain speaker-related errors. More examples are in Appendix A.1.

Based on this perspective, we have performed human evaluation[1] on the summaries generated by BART$_{base}$ model to figure out whether they adequately include the contents of the conversation. The results showed that only 47% of the samples can be regarded as correct summaries. The rest mainly contain incorrect contents w.r.t. references, reasoning, and gendered pronouns. One interesting finding is that nearly half of the incorrect summaries have speaker-related errors. As shown in Table 1, even though the generated summaries are well-constructed and seem plausible, they are obviously wrong since they describe participants' actions with incorrect speakers. Specifically, the speakers in D1 (i.e., *Betty* and *Isabella*) should be replaced, and one of the speakers in D2 (i.e., *Anna*) should be removed to make the draft summaries correct.

To address the aforementioned finding, this paper mainly focuses on improving the quality of the dialogue summary in terms of correcting speakers. Existing works proposed post-editing methods for abstractive text summarization (Cao et al.,

*These authors equally contributed to this work.

[1]We choose 100 test set samples provided by Chen and Yang (2020).

2020; Dong et al., 2020) and table-to-text (Iso et al., 2020), but they mainly focus on correcting summary of the general corpora (e.g., news articles) or facts in a knowledge base, which are somewhat different from the multi-party conversation. Some researches for dialogue summarization (Zhao et al., 2020, 2021) utilized utterance-level representations by constructing dialogue graph, but they lack in leveraging speaker information explicitly.

In this work, we propose a speaker-focused post-correction model for abstractive dialog summarization. We first construct the dataset by using self-supervised speaker manipulation strategies, which corrupt the speakers in summary on purpose. During training, our model predicts whether the speakers are corrupted or not by using the speaker correction type discriminator and then generates a corrected summary according to the correction type via the speaker-focused correction generator.

Our main contributions are as follows. 1) We show that the existing dialogue summarization model often generates incorrect summaries that contain speaker-related errors (i.e., insertion, deletion, and replacement) through human evaluation. Based on this, we design self-supervised speaker manipulation strategies to construct the post-editing data without extra annotations. 2) We propose the highly effective speaker-focused post-correction model not only to capture speaker incorrectness but also adequately revise the draft summary. To the best of our knowledge, it is the first attempt to adopt the post-editing method w.r.t. the speakers in abstractive dialogue summarization. 3) Experimental results show that the revised summaries are significantly improved compared to the draft summaries in both quantitative and qualitative evaluations.

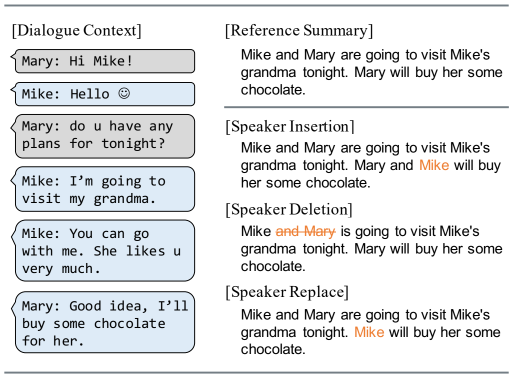


Figure 1: A corruption example with the speaker manipulation strategies. Words in orange represent modified speakers.

2 Proposed Approach

2.1 Problem Definition

Given a dialogue context $D = \{w_1, w_2, ..., w_n\}$, where n denotes the number of tokens, an abstractive summarization model aims to generate a draft summary $Y^d = \{w_1, w_2, ..., w_m\}$, which is conditioned on the likelihood of $p(Y^d|D)$. However, a draft summary might contain incorrect speaker information, which is caused by the conditional maximum-likelihood objective (Li et al., 2018).

Our proposed model generates a corrected summary $Y^c = \{w_1, w_2, ..., w_k\}$ as follows. First, we corrupt a reference summary Y^r using the self-supervised speaker manipulation strategies to obtain a corrupted summary Y^s. Second, given the dialogue context D and a draft summary Y^d, the required speaker correction type C is predicted, which can be formulated as $p(C|D, Y^d)$. Finally, the speaker-focused correction generator is trained to maximize the conditional distribution of $p(Y^c|D, Y^d, C)$. During training, we use either Y^s or Y^r as input summary and train the model to recover them to Y^r (i.e., Both corrupted and uncorrupted summaries are utilized to prevent over-correction (Section 2.2)).

2.2 Data Creation with Self-Supervised Speaker Manipulation Strategies

Given a reference summary Y^r, we obtain a corrupted summary Y^s by conducting the self-supervised speaker manipulation strategies: speaker *insertion*, *deletion*, and *replacement*. Figure 1 represents an example of the proposed strategies. First, we extract a list of the speakers from the dialogue context. Second, we randomly choose any speaker to be corrupted and apply speaker insertion, deletion, or replacement functions at a random rate. For the *speaker insertion*, we arbitrarily select a speaker and add it to another speaker with a conjunction or comma. Likewise, for the *speaker deletion*, we remove a speaker followed by a comma and conjunction with other speakers. In the case of *speaker replacement*, we randomly choose a speaker and replace it with another speaker. We also adjust the subject-verb agreement using heuristics as the number of speakers change.

Finally, we label the required correction type according to the speaker manipulation function that is used. For example, if the *speaker insertion* is conducted on a reference summary, we label the required correction type as *deletion*. The required correction type label is used to train the speaker correction type discriminator in Section 2.3. Among the training set, we set the ratio of uncorrupted and corrupted examples to 1:1 to prevent over-correction (Section 2.4). The whole procedure of the speaker manipulation strategies is described in Appendix A.3.

2.3 Speaker Correction Type Discriminator

We utilize the BART$_{large}$ encoder-decoder (Lewis et al., 2020) to discriminate which type of speaker correction is required on a draft summary. Given a dialogue context D and a draft summary Y^d, our speaker correction type discriminator (SCTD) aims to predict required correction type C, where $C \in \{NO, INS, DEL, REP\}$. Each correction type denotes *no needs to be changed*, *the speaker needs to be inserted, deleted,* and *replaced*, respectively.

The input to the BART is a concatenation of a dialogue context D and a draft summary Y^d, which is represented as $[\textit{<BOS>}, D, \textit{<EOS>}, Y^d, \textit{<EOS>}]$. Then, the output representation of the last <*EOS*> token $h_{\textit{<EOS>}} \in \mathbb{R}^{d_h}$, where d_h denotes a size of output representation, is used to classify the required correction type. We utilize a single-layer feed-forward neural network (FFNN), denoted as,

$$\begin{aligned} Z &= (W_1 h_{\textit{<EOS>}} + b_1) \\ \hat{C} &= \text{softmax}(W_2 Z + b_2), \end{aligned} \tag{1}$$

where $W_1 \in \mathbb{R}^{d_h \times d_h}$ and $W_2 \in \mathbb{R}^{4 \times d_h}$ are trainable parameters. The parameters of the shared BART are represented as Θ_{shd} and those of a single-layer FFNN are represented as Θ_{disc}. The objective is minimizing the negative log-likelihood (NLL) loss: $\mathcal{L}_{\text{SCTD}}(\Theta_{shd}, \Theta_{disc}) = -\sum \log p(C|D, Y^d)$. Another objective of the SCTD is to impose interpretability to the draft summary, which leads to preventing the SCG (Section 2.4) from making a false-positive correction.

2.4 Speaker-focused Correction Generator

Speaker-focused Correction Generator (SCG) utilize the shared BART$_{large}$ to generate a speaker-focused corrected summary. Given a dialogue context D, a draft summary Y^d, and a required correction type C, the input to the BART is represented as $[\textit{<BOS>}, \textit{<COR>}, D, \textit{<EOS>}, Y^d, \textit{<EOS>}]$. Here, we construct the special correction token $\textit{<COR>} \in \{\textit{<NO>}, \textit{<INS>}, \textit{<DEL>}, \textit{<REP>}\}$, which is predicted by SCTD. In this manner, the SCG conditionally generates a corrected summary based on the required speaker correction type. We optimize the model by minimizing the NLL loss: $\mathcal{L}_{\text{SCG}}(\Theta_{shd}, \Theta_{gen\_cor}) = -\sum \log p(Y^c|D, Y^d, C)$.

2.5 Speaker Generator

To make the model more robust, we devise an auxiliary task of generating speakers who appeared in the reference summary. Given a dialogue context D and a draft summary Y^d, speaker generator constructs the list of speakers S, where speakers S appeared in the reference summary. We optimize the model by minimizing the NLL loss: $\mathcal{L}_{\text{SG}}(\Theta_{shd}, \Theta_{gen\_spe}) = -\sum \log p(S|D, Y^d)$. This gives an inductive bias to explicitly generates the list of speakers, which guides the SCG to generate more accurate eventually. Note that we utilize this task as an auxiliary task only in training time.

2.6 Joint Learning Procedure

All the proposed tasks are jointly trained, and the final objective is defined as, $\mathcal{L} = \mathcal{L}_{\text{SCTD}} + \mathcal{L}_{\text{SCG}} + \mathcal{L}_{\text{SG}}$. Note that the shared parameters Θ_s are optimized for all tasks.

3 Experiments

3.1 Dataset

We evaluate our proposed methods on the SAMSUM (Gliwa et al., 2019) dialogue summarization dataset. The SAMSUM dataset is a recently proposed English dataset regarding real-life messenger conversations such as chit-chats, meetings, politics, etc. The dataset consists of 14,732, 818, and 819 dialogue–summary pairs for training, validation, and testing, respectively.

3.2 Quantitative Results

We evaluate our proposed model on the test set by using the standard ROUGE (Lin and Och, 2004) metric. For the draft summarization models, we choose BART$_{base}$ and BART$_{large}$, which are powerful baselines for abstractive dialogue summarization. In Table 2, we compare the ROUGE scores of the draft summaries and those of corrected summaries. Overall, the corrected summaries show significantly higher ROUGE-2 and ROUGE-L scores than those of the draft summaries. Specifically, our correction model shows significant improvements

Draft Model	Speaker Generator	Correction Rate (%)	ROUGE-1		ROUGE-2		ROUGE-L	
			Draft	Corrected	Draft	Corrected	Draft	Corrected
BART$_{base}$	✗	9.8	0.488	**0.493**	0.234	**0.261**	0.447	**0.460**
	✓	9.5	**0.477**	0.473	0.225	**0.251**	0.434	**0.437**
BART$_{large}$	✗	5.4	0.472	**0.475**	0.213	**0.233**	0.428	**0.442**
	✓	3.9	**0.454**	0.444	0.186	**0.194**	0.405	**0.417**

Table 2: ROUGE scores on the test set. “Correction Rate" indicates the rate of the corrections that have been conducted by the model.

Draft Model	Speaker Generator	Correction Rate (%)	After Correction		
			Better	Worse	Same
BART$_{base}$	✗	9.8	60%	21%	19%
	✓	9.5	**61%**	19%	20%
BART$_{large}$	✗	5.4	47%	24%	29%
	✓	3.9	**54%**	26%	20%

Table 3: Human Evaluation results on the test set.

Speaker Generator	F1-Score				
	NO	*INS*	*DEL*	*REP*	Micro AVG
✗	94.67	87.30	95.40	**89.44**	93.03
✓	**95.23**	**92.91**	**96.47**	89.27	**93.89**

Table 4: Automatic Evaluation of the SCTD for each correction type.

Speaker Generator	F1-Score				
	NO	*INS*	*DEL*	*REP*	Micro AVG
✗	98.36	93.10	95.08	96.67	95.83
✓	**100.0**	**96.55**	**98.36**	**98.36**	**98.32**

Table 5: Human Evaluation of the SCG w.r.t. the conditional generation for each correction type.

in ROUGE-2 on BART$_{base}$ draft model (absolute improvements of 2.7%).

3.3 Human Evaluation

We also conduct a human evaluation to validate the corrected summaries generated by our proposed model. Given a dialogue context, reference summary, draft summary, and corrected summary, we asked five annotators from Amazon Mechanical Turk (AMT) to judge a corrected summary is either better, worse, or same compared to a draft summary. An example given to annotators and more details are described in Appendix A.4. As reported in Table 3, the corrected summaries show significantly better results for both BART$_{base}$ and BART$_{large}$ draft models after corrections. Specifically, the speaker generator has little effect on the model when the draft summaries are generated by BART$_{base}$, but shows a performance improvement when the draft model is BART$_{large}$ (Better: 47% $\rightarrow$ 54%). The reason why the ratio of better results is lower in BART$_{large}$ compared to that of BART$_{base}$ is that the BART$_{large}$ draft model mostly produces more complete summaries than BART$_{base}$ with fewer speaker errors.

3.4 Conditional Generation Analysis

In this analysis, we verify the performance of SCTD and how SCG conditionally generates a corrected summary based on the predicted speaker correction type.

For the SCTD evaluation, we corrupt a reference summary following Section 2.2 and use a corrupted summary as a draft summary. Then, the SCTD predicts which type of speaker correction is required on the draft summary. As reported in Table 4, when utilizing the speaker generator objective as an auxiliary task, the SCTD shows higher F1 scores for all correction types except *REP*. We also observe that the SCTD with speaker generator shows 95.23 in F1 score for *NO* label. This result leads to prevent the SCG from producing false-positive corrections while saving the amount of computation since the summary that predicted as *NO* label is not corrected.

For the SCG evaluation, we sampled 120 (30 for each of four correction type) examples and asked four annotators to judge which operation (*e.g., NO*, *INS*, *DEL*, *REP*) is actually performed when generating a corrected summary given the speaker correction type from SCTD and the draft summary. Then, we measure how well the predicted and actual correction types align using the F1-score. From Table 5, we observe that the SCG with speaker generator shows higher F1 scores for every correction type (98.32% on average). This suggests that our SCG can conditionally generate a well-corrected summary based on the required speaker correction type.

4 Conclusion

In this paper, we pointed out that current dialogue summarization models have problems summarizing the multi-party conversation. To address these problems, we proposed the speaker-focused post-correction model, which can be applied to any abstractive dialogue summarization model. Experimental results show that our model adequately corrects a draft summary.

Acknowledgements

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2018-0-01405) supervised by the IITP (Institute for Information Communications Technology Planning Evaluation).

This work was supported by Institute for 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

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258.

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.

Jiaao Chen and Diyi Yang. 2021. Structure-aware abstractive conversation summarization via discourse and action graphs. *arXiv preprint arXiv:2104.08400*.

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. In *Advances in Neural Information Processing Systems*, volume 32.

Yue Dong, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu. 2020. Multi-fact correction in abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9320–9331.

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.

Hayate Iso, Chao Qiao, and Hang Li. 2020. Fact-based Text Editing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 171–182.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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.

Haoran Li, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2018. Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive sentence summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1430–1441.

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.

Valentin Malykh, Konstantin Chernis, Ekaterina Artemova, and Irina Piontkovskaya. 2020. SumTitles: a summarization dataset with low extractiveness. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5718–5730.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8026–8037.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Chien-Sheng Wu, Linqing Liu, Wenhao Liu, Pontus Stenetorp, and Caiming Xiong. 2021. Cordial: Coarse-to-fine abstractive dialogue summarization with controllable granularity.

Lulu Zhao, Weiran Xu, and Jun Guo. 2020. Improving abstractive dialogue summarization with graph structures and topic words. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 437–449.

Lulu Zhao, Zeyuan Yang, Weiran Xu, Sheng Gao, and Jun Guo. 2021. Improving abstractive dialogue summarization with conversational structure and factual knowledge.

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*.

A Appendix

Dialogue Index: 116
luke: Hey, was just thinking, we should avail ourselves for team selection tomorrow regardless of our injuries martial: thats what i was thinking also luke: we should let Jose know that tomorrow martial: the first thing in the morning infact luke: the fixtures are really piling up and we need to help the team martial: yeah, thats for sure, we are a family luke: we will the coach know that we are ready to play martial: despite the little pain, me i'm ready luke: me too martial: so we meet up at carrington and go to his office very early luke: yeah, both of us martial: ok, we'll go together luke: cool martial: ok **Draft Summary (BART$_{base}$)** Lukeke, Martial and Jose are going to meet at Carrington and go to the coach's office very early tomorrow. ⇒ Jose is the coach. ↪ Deletion
Dialogue Index: 158
Dave: Hey, is Nicky still at your place? Her phone is off Sam: She just left Dave: Thanks! **Draft Summary (BART$_{base}$)** *Nicky left Dave's place and her phone is off.* ⇒ Nicky left Sam's place. ↪ Replacement

Table 6: Examples of the incorrect summaries that contain speaker-related errors. Dialogue index denotes the index of test set. All the indices are provided by Chen and Yang (2020). ⇒ represents the explanations why the summary is incorrect and ↪ represents the required speaker correction type.

A.1 Draft Summary Evaluation

We describe more examples of the draft summary evaluation in Table 6. They are all generated by BART$_{base}$, and we focus on analyzing the examples that contain speaker-related errors.

A.2 Implementation Details

We implemented our model using the PyTorch (Paszke et al., 2019) library. For the BART-based correction model, we adopt the pre-trained language model BART$_{large}$ based on the hugging face open source[2] (Wolf et al., 2020). For fine-tuning, we trained the correction model using Adam optimizer (Kingma and Ba, 2014) with a batch size of 32 and an initial learning rate of 3e-05. We also utilized the pre-trained BART$_{base}$ and BART$_{large}$ as the draft summarization models. We trained both models using Adam optimizer with a batch size of 32 and an initial learning rate of 3e-05. The correction model is trained for 5 epochs, and BART$_{base}$ and BART$_{large}$ based draft models are trained for 8 epochs and 4 epochs, respectively, showing the best performance on the validation set. The average runtime of each epoch was about 20 minutes. All experiments were conducted with 4 Tesla V100 GPUs. Our code is publicly available[3].

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

Inputs:
D - dialogue context
Y^r - reference summary
$FLAG$- corruption flag
Outputs:
Y^s- corrupted summary
C- required correction type

```
function SPEAKER_MANIPULATION(D, Y^r, FLAG)
    if FLAG then
        speaker_list ← EXTRACT_SPEAKERS(D)
        func_type ← random.choices[ins, del, rep]
        if func_type == ins then
            Y^s ← SPEAKER_INS(Y^r, speaker_list)
            C ← del
        else if func_type == del then
            Y^s ← SPEAKER_DEL(Y^r, speaker_list)
            C ← ins
        else if func_type == rep then
            Y^s ← SPEAKER_REP(Y^r, speaker_list)
            C ← rep
        end if
    else
        Y^s ← Y^r
        C ← no
    end if
    return Y^s, C
end function
```

Table 7: Procedure to create dataset with self-supervised speaker manipulation strategies.

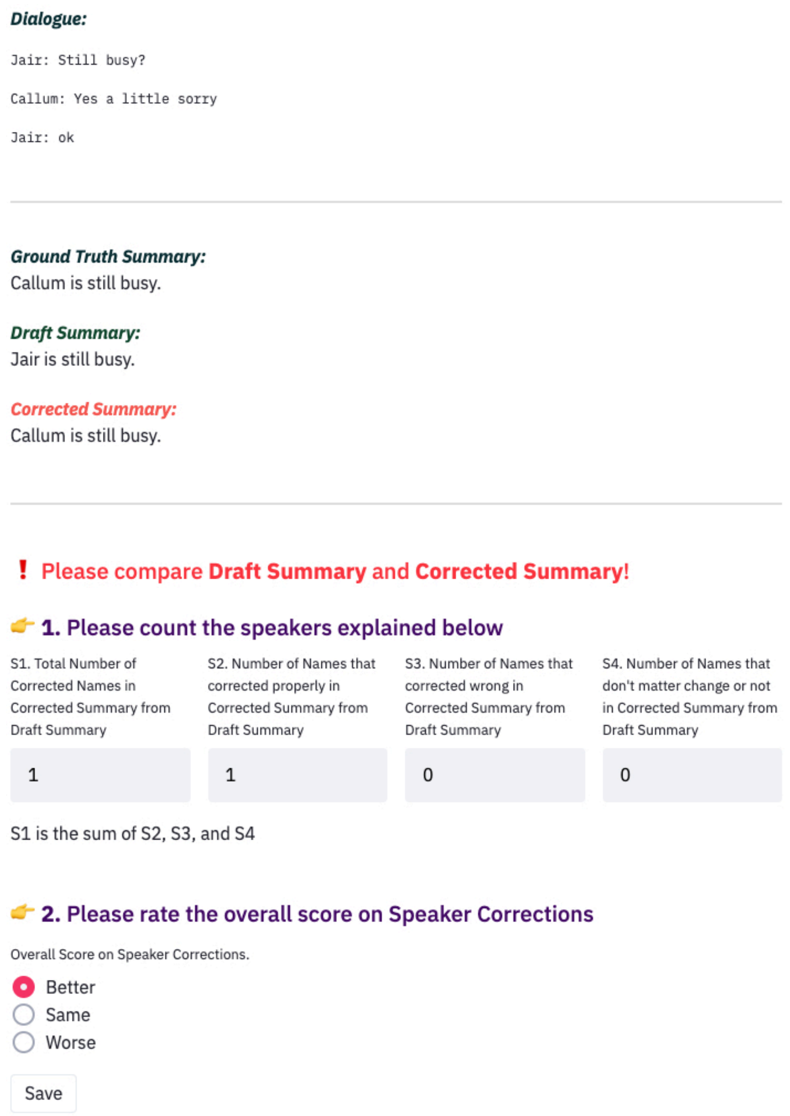
A.3 An algorithm of Speaker Manipulation Strategies

Table 7 represents the procedure of data creation with our self-supervised speaker manipulation strategies. Here, we decide whether or not to corrupt the reference summary through FLAG.

A.4 Annotations for Human Evaluation

We first showed Turkers a draft summary and corrected summary by our models. In order to focus on the evaluation of speaker corrections, we asked Turkers to count the number of speakers that changed appropriately, badly, or the same as in Figure 2. By counting the number of speakers, the overall assessment of the speaker corrections was

[3] Github repository will be available upon paper acceptance.


Dialogue:

```
Jair: Still busy?

Callum: Yes a little sorry

Jair: ok
```

Ground Truth Summary:
Callum is still busy.

Draft Summary:
Jair is still busy.

Corrected Summary:
Callum is still busy.

! Please compare **Draft Summary** and **Corrected Summary**!

1. Please count the speakers explained below

S1. Total Number of Corrected Names in Corrected Summary from Draft Summary	S2. Number of Names that corrected properly in Corrected Summary from Draft Summary	S3. Number of Names that corrected wrong in Corrected Summary from Draft Summary	S4. Number of Names that don't matter change or not in Corrected Summary from Draft Summary
1	1	0	0

S1 is the sum of S2, S3, and S4

2. Please rate the overall score on Speaker Corrections

Overall Score on Speaker Corrections.

- Better
- Same
- Worse

Save

Figure 2: An example given to AMT annotators.

evaluated with Turkers' objectivity. The average Fleiss' Kappa represents moderate level of inter-rater agreement.

A.5 Qualitative Analysis

We also conduct qualitative analysis w.r.t each correction type (i.e., speaker insertion, deletion, and replacement). As illustrated in Table 8, our speaker-focused post-correction model adequately corrects draft summaries for all correction types.

Correction Type	Examples	
Insertion	**Dialogue** Andy: Hi nephew! Paul: Hi uncle! Andy: Are you home? I'm nearby and thought I would drink coffee with you :) Paul: Yup. I'm home. Feel free to come! Andy: If that is ok I will visit you in about 1 hour. Paul: Sure. A lot of political cases for us to talk about :D Andy: Haha. No. Andy: Too much politics with Hannah's father. Andy: I have enough arguments over politics forever. Paul: Hahah. Ok. Waiting for you then. Andy: See you.	**Ground Truth Summary** Andy is going to visit Paul in about 1 hour. **Draft Summary (BART$_{base}$)** Andy will meet Paul for coffee in 1 hour. Andy has a lot of political issues to discuss. **Corrected Summary** Andy will meet Paul for coffee in 1 hour. Paul and Andy have a lot of political issues to discuss.
Deletion	**Dialogue** Julia: Hey, what time are you getting home? Bert: 8-ish. Why? Julia: I was wondering if we should wait for you with the dinner? Bert: Yeah, that would be nice of you. I'll try to get there on time Julia: Ok. Call me if you're running late Bert: I will. xx	**Ground Truth Summary** Julia will be waiting for Bert with the dinner. Bert is coming home around 8. **Draft Summary (BART$_{base}$)** Julia and Bert will wait for Bert with dinner. **Corrected Summary** Julia will wait for Bert with dinner.
	Dialogue Bradley: haha look a cat invaded the pitch at Goodison <file_other> Jill: hahahaha Julia: what a sweet little black ball of fur Jill: here's the video :D <file_other> Julia: haha Bradley: and the commentary :D Bradley: that's the best entertainment Everton fans have had all season :D	**Ground Truth Summary** A sweet little black cat got into the pitch during the Everton's football match. **Draft Summary (BART$_{base}$)** Bradley, Jill, Julia and Julia are talking about the football match at Goodison. **Corrected Summary** Bradley, Jill and Julia are talking about the football match at Goodison.
Replacement	**Dialogue** Randolph: Honey Randolph: Are you still in the pharmacy? Maya: Yes Randolph: Buy me some earplugs please Maya: How many pairs? Randolph: 4 or 5 packs Maya: I'll get you 5 Randolph: Thanks darling	**Ground Truth Summary** Maya will buy 5 packs of earplugs for Randolph at the pharmacy. **Draft Summary (BART$_{base}$)** Randolph will buy 5 pairs of earplugs for Maya. **Corrected Summary** Maya will buy 5 pairs of earplugs for Randolph.
	Dialogue Paula: Why do they make this game with super hard levels? Stew: No idea. I hate those. Paula: It really makes it not fun at all. Stew: Yep. Paula: I just can get past 637 no matter what I do. Stew: Did you try looking up the cheats online? Paula: Brilliant!	**Ground Truth Summary** Paula cannot get past level 637 in her game. She will look up the cheats online. **Draft Summary (BART$_{base}$)** Paula hates the game with super hard levels. Stewart tries looking up the cheats online. **Corrected Summary** Stew hates the game with super hard levels. Paula tries looking up the cheats online.
	Dialogue Willy: Your car is friggin' awesome!! Vinny: I know ;) No, but seriously, I've always wanted a Mustang, and a red one too! Willy: Maybe you can lend it to me for a day or so :) Vinny: Yeah, right. We can car pool together a couple of days a week. Willy: Ok, deal.	**Ground Truth Summary** Willy and Vinny will car pool with Winny's red Mustang. **Draft Summary (BART$_{base}$)** Willy will lend his car to Ginny for a day or so. They will car pool together a couple of days a week. **Corrected Summary** Vinny will lend his car to Will for a day or so. They will car pool together a couple of days a week.
	Dialogue Jair: Still busy? Callum: Yes a little sorry Jair: ok	**Ground Truth Summary** Callum is still busy. **Draft Summary (BART$_{base}$)** Jair is still busy. **Corrected Summary** Callum is still busy.

Table 8: Qualitative analysis w.r.t each correction type (i.e., Insertion, Deletion, and Replacement). Words in red represent the incorrect speakers that should be corrected and words in blue represent the correction results.