



Negative Guided Abstractive Dialogue Summarization

Junpeng Liu^{*, \diamond 1}, Yanyan Zou ^{\diamond 2}, Yuxuan Xi^{*,3}, Shengjie Li², Mian Ma², Zhuoye Ding², Bo Long²

¹Beijing University of Posts and Telecommunications, Beijing, China

²JD.com, Beijing, China

³Peking University, Beijing, China

jeepliu@bupt.edu.cn, xiyuxuan@pku.edu.cn,

{zouyanyan6, lishengjie1, mamian, dingzhuoye, bo.long}@jd.com

Abstract

The goal of the abstractive dialogue summarization task is to generate a shorter form of a long conversation while retaining its most salient information, which plays an important role in speech. Unlike the well-structured text, such as scientific articles and news, dialogues often comprise of utterances coming from multiple interlocutors, where the conversations are often informal, verbose, repetitive, and sprinkled with false-starts, backchanneling, reconfirmations, hesitations as well as speaker interruptions, which might introduce much noisy information and thus brings new challenges of summarizing dialogues. In this work, we extend the widely-used sequence-to-sequence summarization framework with a negative guided mechanism, which allows models to explicitly perceive the unnecessary pieces (i.e., noise) of a dialogue and thus focus more on the salient information. Specifically, the negative guided mechanism has two main components, negative example construction and negative guided loss. We explore two different ways to constructing the negative examples and further calculate the negative loss. Extensive experiments on the benchmark datasets demonstrate that our method significantly outperforms the baselines with regard to both semantic matching and factual consistent based metrics. We also elicit the human efforts to prove the performance gains.

Index Terms: natural language generation, speech summarization, abstractive dialogue summarization

1. Introduction

Dialogue summarization aims to condense a long dialogue into its shorter form with its most informative content preserved while the unnecessary pieces are ignored, which allows applications in speech to quickly grasp the key information of a dialogue comprising of utterances coming from multiple interlocutors. Generally, the task of text summarization can be typically classified into two categories: extractive and abstractive summarization. The former [1] aims to select important sentences from the input text and concatenate such extracted ones as a summary. Differently, the abstractive summarization task [2] rewrites the source text and generates a summary that may contain novel words and phrases not featured in the input. For the typical document summarization problem, such as well-structured scientific articles and news, both the extractive and abstractive fashions perform well. Differently, a conversation often consists of utterances coming from two or more interlocutors, where the utterances are often informal, verbose, and repetitive, and sprinkled with false-starts, backchanneling, reconfirmations, hesitations, as well as speaker interruptions [3]. Moreover, the salient information

*refers to work done during internship at JD.com.

\diamond means equal contributions. Corresponding to Yanyan Zou.

Samantha:	How are you doing today?
Robyn:	<u>better, but I really exaggerated with beer last night</u>
Samantha:	was it really only beer?
Robyn:	sure, why?
Samantha:	I don't know, your eyes, behaviour
Robyn:	you want to say I took drugs?
Samantha:	I'm only asking, I don't have anything against drugs
Robyn:	but I have and I never take them
Samantha:	ok, sorry, I didn't intend to offend you
Robyn:	<u>I just drank too much, that's it</u>
Samantha:	<u>much too much</u>
Robyn:	<u>yes, I lost control a bit. I am sorry for that</u>
Samantha:	<u>important that you feel better today</u>

Summary:	Robyn drank too much beer last night and lost control. She is feeling better today.
----------	---

Figure 1: An example of dialogue with its summary. Underline: salient utterances

of a long dialogue is often scattered throughout the whole conversation. Therefore, it is challenging and impractical to perform extractive summarization on the dialogues [4]. As a result, in this work, we focus on the abstractive summarization approach of dialogues, aiming at automatically generating a succinct summary with most salient information retained, exemplified by a dialogue-summary instance in Figure 1.

One straightforward solution to summarizing conversations is to directly adopt existing summarization systems designed for well-structured text [5, 6] or to employ hierarchical models to capture features from different turns of different interlocutors [7, 8]. Most of such systems are deployed based on a sequence-to-sequence architecture with attention mechanism in recent literature [9, 10], where all the source input are taken as input by an encoder, and then output of the encoder is fed into a decoder to generate the final summary. Despite the attentive sequence-to-sequence models have proved effective in various text generation tasks, like neural machine translation [9, 10], its performance in abstractive summarization still needs further improvements. Since the natural language is complex, the source input of a summarization model often contains redundant information, i.e., noises, especially for the dialogue summarization [11]. Different from the neural machine translation task where all information of the source input is required to be translated and retained in the target output, the abstractive summarization task only needs to keep the salient information while the unnecessary pieces can be ignored. However, the attentive sequence-to-sequence model simply takes as input all the dialogue words, which might introduce noises in the generated summary.

To alleviate such an issue, recent studies [12] derived a global vector based on the output of the encoder, to guide the noise filtering process, where CNNs are often used to filter out

the noisy information. However, such a line of noise filtering methods might not be a good choice with recent new pretraining-finetuning paradigms, requiring extra networks apart from the original sequence-to-sequence architecture. In this work, we propose to extend the sequence-to-sequence framework with a negative guided mechanism, which allows a summarization model to explicitly perceive the noises of a source input and then thus pay more attention to the salient information. Specifically, the negative guided mechanism has two main components, i.e., negative example construction and negative guided loss. For the negative example construction, we utilize the ROUGE-1 recall score [13] to select the salient utterances from the input dialogue, and the rest is considered as redundant ones, based on which we construct the negative examples. Then a negative guided loss is designed to cooperate with the sequence-to-sequence likelihood loss to guide the model to perceive the unnecessary pieces as well as to focus more on the salient utterances. Such a mechanism is pluggable and thus can be used to equip any sequence-to-sequence framework with no extra networks required. We conducted extensive experiments on the benchmark datasets which demonstrates the proposed method significantly outperforms the baselines with regard to both semantic matching and factual consistent based metrics. The human evaluation also proves the effectiveness.

2. Method

2.1. Problem Formulation

We regard the abstractive dialogue summarization task as a sequence-to-sequence learning problem. The Transformer [9] is adopted as our backbone architecture, where the model encoder takes as input the whole dialogue utterances and the decoder generates a corresponding summary. To be specific, for a dialogue $D = (u_1, u_2, \dots, u_{|D|})$, consisting of $|D|$ utterances, associated with its corresponding summary $Y = (y_1, y_2, \dots, y_{|Y|})$ in the length of $|Y|$, the goal is to learn the optimal model parameters θ and to estimate the conditional probability:

$$P(Y|D; \theta) = \prod_{i=1}^{|Y|} p(y_i | y_{1:i-1}, D; \theta)$$

where $y_{1:i-1}$ denotes the first $i-1$ tokens of the output sequence (i.e., $y_{1:i-1} = (y_1, y_2, \dots, y_{i-1})$). Given the training set $(\mathcal{D}, \mathcal{Y})$, a model is trained to maximize the log-likelihood by minimizing:

$$\mathcal{L}_{MLE}(\theta; \mathcal{D}, \mathcal{Y}) = - \sum_{(D, Y) \in (\mathcal{D}, \mathcal{Y})} \log P(Y|D; \theta)$$

2.2. Negative Guided Mechanism

An utterance is defined as a *salient utterance* when the ROUGE-1 recall score [13] between it and the gold summary is larger than zero, while an utterance is considered as a *redundant utterance* where the ROUGE-1 recall score is zero. For example, the utterances marked with underlines in Figure 1 are salient ones.

We design two settings to introduce negative signals:

REDUNDANT: We feed all the redundant utterances of an input dialogue into an already trained dialogue summarization model, and consider the generated summary Y' as the negative output sequence for the original input utterance D . The negative loss is then calculated as:

$$\mathcal{L}_{NEG}^{\text{REDUNDANT}}(\theta; \mathcal{D}, \mathcal{Y}') = - \sum_{(D, Y') \in (\mathcal{D}, \mathcal{Y}')} \log(1 - P(Y'|D; \theta))$$

which is similar to the sequence-level unlikelihood objective [14].

NULL: Removing all the salient utterances from a dialogue D results in an input sequence D' of a negative example. Here, we simply regard a null sequence Y' as the output and the loss is:

$$\mathcal{L}_{NEG}^{\text{NULL}}(\theta; \mathcal{D}', \mathcal{Y}') = - \sum_{(D', Y') \in (\mathcal{D}', \mathcal{Y}')} \log P(Y'|D'; \theta)$$

The final loss is then calculated as:

$$\mathcal{L} = \mathcal{L}_{MLE} + \alpha \mathcal{L}_{NEG}^{\beta}$$

where α is a contribution weight for the negative guided loss, considered as a hyperparameter, and β refers to either REDUNDANT or NULL.

The proposed negative guided objective, acting as an auxiliary task, can contribute to the primary dialogue summarization task during training phase. The goal is to minimize the loss \mathcal{L} , i.e., maximizing the probability of generating the summary Y given the original dialogue D , while minimizing the probability of producing Y' given the original dialogue D (i.e., REDUNDANT) or maximizing the probability of producing Y' given negative input sequence (i.e., NULL) which guides the model to pay less attention to the negative signals while more eyes on the salient parts.

3. Experiments

3.1. Datasets

To evaluate the effectiveness of our proposed method and compare the approach with other baselines, we conduct experiments on the widely-adopted abstractive dialogue summarization dataset, i.e., the SAMSum corpus [6]. Such a corpus consists of natural message-like chat dialogues expressed in English written by linguists fluent in English, each of which is annotated with a summary created by language experts. Each dialogue-summary instance follows the same specified format that a colon at the beginning of an utterance separates the interlocutor's name and their speaking content¹. The whole corpus are divided into three splits. The training set consists of 14,732 dialogue-summary pairs, while the validation and test set contain 818 and 819 instances individually. We list the detailed data statistics of each split (i.e., training, validation, and test) with regard to average speakers, utterances and words in Table 1.

3.2. Implementation Details

We adopted the sequence-to-sequence Transformer model as our backbone architecture, which is implemented using Fairseq toolkit. To be specific, our model is initialized with a pre-trained sequence-to-sequence, i.e., BART [10]. Thus they share the same architectures, a 12-layer encoder-decoder Transformer. Each layer has 16 attention heads, and the hidden size and feed-forward filter size is 1024 and 4096, respectively, resulting in 400M trainable parameters. The dropout rates for all layers are set to 0.1. The optimizer is Adam [16] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The peak learning rates for all experiments are set to $3e-5$ with 200 warmup steps. We also adopted the same learning rate schedule strategies as in [9]. The maximum number of tokens in

¹We note that it is likely to improve the model performance by enriching the vocabulary with frequently appeared emojis in the training split in order to avoid too many [UNK] in the input sequence [15]. However, to make fair comparisons with other baselines, we simply adopted the original version of the vocabulary for the pre-trained model.

each batch is 800. The base dialogue summarization model used to generate negative sequence for REDUNDANT is first trained on the whole original training set with only likelihood loss for 2 epochs. Then, the model is further trained with both likelihood and negative guided loss for 5 epochs with early stops. Each epoch takes around 0.7 hours on single Tesla P40 GPU. The contribution weight α is set to 0.05 for REDUNDANT and 0.001 for NULL individually. All hyperparameters are set based on the performance of the validation set.

3.3. Baseline

We adopted several previous state-of-the-art abstractive dialogue summarization models as the baseline systems:

- Lead3 is a widely-used method in the extractive document summarization task, which directly takes the first three leading sentences of an input text as its summary. In the dialogue scenario, we also take the first leading three utterances as the summary while the speaker’s name at the beginning of each utterance is removed.
- PTGen [2] equips a sequence-to-sequence generation model with the copy and coverage mechanisms that allows to copy words from the source input.
- FastAbs-RL [17] first selects salient sentences and then employs the reinforcement learning with sentence-level policy gradient methods to rewrite the selected sentences as generated summary.
- DynamicConv + GPT-2/News [18] proposes a lightweight dynamic convolutions to replace the self-attention modules in the Transformer layers. Following [19], we also consider two variants. DynamicConv + GPT-2 uses the GPT-2 [20] to initialize the token embeddings. DynamicConv + News considers news summarization corpus CNN DailyMail [21] as extra training data.
- BART [10] is a pre-trained sequence-to-sequence Transformer model, with two versions BART_{BASE} and BART_{LARGE}. In this work, we adopts the BART_{LARGE}.

3.4. Automatic Evaluation

The full-length F1-based ROUGE score [13] is often used as the automatic evaluation metric to measure the quality of summary output generated by different systems, which calculates the similarity between the predicted summary and the ground truth via string match. To evaluate the effectiveness of the proposed model and compare it with other baselines, we also used the full-length F1-based ROUGE score as a measure metric. It is worth noting that the ROUGE scores might vary with different ROUGE toolkits. To be specific, we adopted the files2rouge² package based on the official ROUGE-1.5.5.perl script to get the full-length ROUGE-1, ROUGE-2 and ROUGE-L F-measure scores. The recent popular automatic evaluation metric for text generation, BERTSCORE [22], is also considered as a complementary for model comparisons³.

The above metrics mainly focus on the semantic similarity between the system output and the gold summary, based on either string match (e.g., ROUGE) or meaning similarity (e.g., BERTSCORE). Furthermore, we also consider the QuestEval [23] to evaluate the factual consistency between the source

input and the generated summary. To be specific, given a dialogue content and a summary, QuestEval first extracts question answers (e.g., named entities) from either the input text or the summary, and then generates natural language questions from the input text or the summary correspondingly conditioned on the extracted answers. A Question Answering (i.e., QA) model is employed to consume the input text to answer the questions derived from the summary, resulting in a score, denoted as the PRECISION score. PRECISION implies that a summary should contain only factual information consistent to the input text. Similarly, the QA model is also applied to address the questions constructed from the input, producing another score, namely the RECALL score, showing that the summary should contain the most important information from the source text. The final QuestEval score is the harmonic mean of the PRECISION and RECALL, i.e., the F1-measure score. We adopted the version⁴ with learned weights for questions, which has proved high correlation with human judged consistency and relevance [23].

The results are listed in Table 2, where the highest score for each metric is highlighted with bold. We can see that, in terms of the semantic similarity-based metrics, the REDUNDANT and NULL both lead to significant performance gains (according to the ROUGE scripts, $p < 0.05$), demonstrating the effectiveness of negative guided mechanism. With regard to the factual consistency metric QuestEval, the variant with NULL obtained the highest F1 score, which demonstrates its effectiveness to generate the factual consistent summaries. Comparing the REDUNDANT and NULL, the latter is better than the former for most of the metrics. One possible reason is that, the negative output sequence of REDUNDANT is generated by a already trained dialogue summarization system, which is in a deterministic space. However, the NULL treats the null sequence as a negative output sequence, which fits the abstractive summarization task better, since all but the golden summary are negative signals.

3.5. Human Evaluation

In addition to the automatic evaluation, we also elicit feedback from human efforts to evaluate the generated summaries from different summarization systems. We compared our best performing model (i.e. +NULL) with two baselines, PTGen [2] and BART [10], as well as the gold summary label, that is human reference. 100 dialogues are randomly picked from the test split of SAMSum dataset. 10 participants are presented with a dialogue and its paired candidate summaries, including human references (denoted as Gold) and the generated outputs by three models individually. For each selected dialogue, they are asked to rank the candidate output from the best (i.e., 1st) to worst (i.e., 4th) with regard to the four criteria as follows:

- *Fluency*: Is the summary grammatically correct?
- *Informativeness*: Does the summary contains the most informative pieces of the dialogue?
- *Succinctness*: Does the summary express in an abstractive way (e.g., without repetitions or redundant pieces)?
- *Consistency*: Is the summary consistent to the dialogue?

Table 3 listed the proportions of different system rankings and mean rank (lower is better). The output of our proposed method is ranked as the most appropriate summary for 30% of all cases. Overall, we obtain lower mean rank than the other two systems yet still lags behind the Gold one. The Fleiss’ Kappa score [24] among participants is 0.524 that proves fair inter-rater agreement.

²<https://github.com/pltrdy/files2rouge>

³We use version 0.3.8, with default English setting (roberta-large.L17.no-idf.version=0.3.8(hug.trans=4.4.0)-rescaled).

⁴<https://github.com/ThomasScialom/QuestEval>

Table 1: Data statistics of the dialogue summarization dataset, SAMSum, including the total number of dialogues (#Dial.), the total number of utterances (#Utterance), the average number of participants (#Speaker), the average number of turns (#Turns), the average number of words in the dialogue (#Words (Dial.)) and in the summary (#Words (Summary)).

Split	#Dial.	#Utterance	#Speaker	#Turns	#Words (Dial.)	#Words (Summary)
Train	14,732	164,505	2.40	11.17	83.90	20.35
Valid	818	8,860	2.39	10.83	83.26	20.14
Test	819	9,212	2.36	11.25	83.87	20.43

Table 2: Automatic evaluation results on the SAMSum test split. * indicates that the results are significantly different from other baselines in terms of ROUGE scores ($p < 0.05$). PRECISION, RECALL and F1 refer to QuestEval scores, respectively.

Model	Similarity Based				QuestEval		
	ROUGE-1	ROUGE-2	ROUGE-L	BERTSCORE	PRECISION	RECALL	F1
Lead3	31.4	8.7	29.4	-	-	-	-
PTGen	40.1	15.3	36.6	-	-	-	-
DynamicConv + GPT-2	41.8	16.4	37.6	-	-	-	-
FastAbs-RL	42.0	18.1	39.2	-	-	-	-
DynamicConv + News	45.4	20.7	41.5	-	-	-	-
BART	52.6	27.0	42.1	52.1	51.6	28.0	39.8
+ REDUNDANT*	53.4	28.4	43.4	52.9	52.9	28.0	40.4
+ NULL *	53.0	28.9	45.0	53.2	55.3	27.3	41.3

Table 3: Human evaluation on SAMSum: proportions of rankings. MR: mean rank (the lower the better).

Systems	1st	2nd	3rd	4th	MR
PTGen	0.04	0.15	0.32	0.49	3.26
BART	0.21	0.24	0.32	0.23	2.57
Ours	0.30	0.27	0.25	0.18	2.31
Gold	0.45	0.34	0.11	0.10	1.86

4. Related Work

4.1. Document Summarization

Automatic document summarization aims to condense a document into its shorter version where the salient information is retained, often divided into two categories: extractive and abstractive summarization. The extractive summarizer learns to find the informative sentences from the input as its summary, which can be viewed as a sentence ranking problem [25, 26]. The abstractive summarization task learns to generate summaries by rewriting the input, which is a typical sequence-to-sequence learning problem [2, 10, 27]. This work focuses on summarizing dialogues from a sequence-to-sequence learning perspective.

4.2. Dialogue Summarization

The dialogue summarization task is to summarize conversations from multiple interlocutors. [7, 28] designed hierarchical structures to model the conversational utterances from different turns. Recent research progresses utilized the conversational analysis to improve the summarization performance, such as topical information [29, 30, 19, 31], dialogue acts [32] and coreference information [33], which mainly force a summarizer to dig into the salient utterances yet ignoring the redundant parts. Unlikely, this work proposes to utilize the redundant utterances as negative signals to guide the model to focus more on the salient ones.

4.3. Text Generation with Negative Signals

[14] proposed a novel unlikelihood training for text generation, which forces a generation model to assign lower probabilities to unlikely tokens or sequences, such as repetitions and high frequent tokens. [34] applied such a mechanism to the dialogue domain in order to improve the logical consistency of the generated dialogues. [35] proposed a negative training framework to address the malicious and generic responses in dialogue generation models. A base text generation model [14, 34, 35] is designed to identify “bad” or “unlikely” input-output pairs as negative examples. [36] constructed negative examples for text generation tasks by artificially injecting the targeted errors to the ground-truth. This work borrows the idea of negative training to the abstractive summarization task and proposes new negative example construction methods.

5. Conclusion

The abstractive dialogue summarization task plays an important role in speech community. A dialogue usually consists of utterances coming from multiple interlocutors, where the conversation often contains much noisy information, such as repetitions, false-starts, backchanneling, reconfirmations, hesitations, as well as speaker interruptions, which raises new challenges to directly adopt the sequence-to-sequence framework to summarize dialogues. In this work, we propose a negative guided mechanism to make a model explicitly perceive such noise information and thus focus more on the salient utterances. Both automatic and human evaluation have demonstrated the effectiveness of the proposed mechanism. In the future, we plan to explore other approaches to constructing negative examples and other possible solutions to utilizing such negative signals.

6. References

- [1] R. Nallapati, F. Zhai, and B. Zhou, “Summarunner: A recurrent neural network based sequence model for extractive summarization of documents,” in *Proceedings of AAAI*, 2017.

- [2] A. See, P. J. Liu, and C. D. Manning, “Get to the point: Summarization with pointer-generator networks,” in *Proceedings of ACL*, 2017.
- [3] H. Sacks, E. A. Schegloff, and G. Jefferson, “A simplest systematics for the organization of turn taking for conversation,” in *Studies in the organization of conversational interaction*. Elsevier, 1978, pp. 7–55.
- [4] S. Gao, X. Chen, Z. Ren, D. Zhao, and R. Yan, “From standard summarization to new tasks and beyond: Summarization with manifold information,” in *Proceedings of the IJCAI*, 2020.
- [5] G. Shang, W. Ding, Z. Zhang, A. Tixier, P. Meladinos, M. Vazirgiannis, and J.-P. Lorré, “Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted sub-modular maximization,” in *Proceedings of ACL*, 2018.
- [6] B. Gliwa, I. Mochol, M. Biesek, and A. Wawer, “SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization,” in *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, 2019.
- [7] Z. Zhao, H. Pan, C. Fan, Y. Liu, L. Li, M. Yang, and D. Cai, “Abstractive meeting summarization via hierarchical adaptive segmental network learning,” in *Proceedings of WWW*, 2019.
- [8] C. Zhu, R. Xu, M. Zeng, and X. Huang, “A hierarchical network for abstractive meeting summarization with cross-domain pretraining,” in *Findings of the ACL*, 2020.
- [9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proceedings of NeurIPS*, 2017.
- [10] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, “BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” in *Proceedings of ACL*, 2020.
- [11] N. Zhang, S. Deng, Z. Sun, X. Chen, W. Zhang, and H. Chen, “Attention-based capsule networks with dynamic routing for relation extraction,” in *Proceedings of EMNLP*, 2018, pp. 986–992.
- [12] Y. Liu, X. Chen, X. Luo, and K. Q. Zhu, “Reducing repetition in convolutional abstractive summarization,” *Natural Language Engineering*, pp. 1–29, 2021.
- [13] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in *Text Summarization Branches Out*, 2004.
- [14] S. Welleck, I. Kulikov, S. Roller, E. Dinan, K. Cho, and J. Weston, “Neural text generation with unlikelihood training,” in *Proceedings of ICLR*, 2019.
- [15] H. Lee, J. Yun, H. Choi, S. Joe, and Y. L. Gwon, “Enhancing semantic understanding with self-supervised methods for abstractive dialogue summarization,” *Proceedings of Interspeech*, pp. 796–800, 2021.
- [16] D. P. Kingma and J. L. Ba, “Adam: A method for stochastic optimization,” in *Proceedings of ICLR*, 2015.
- [17] Y.-C. Chen and M. Bansal, “Fast abstractive summarization with reinforce-selected sentence rewriting,” in *Proceedings of ACL*, 2018.
- [18] F. Wu, A. Fan, A. Baevski, Y. Dauphin, and M. Auli, “Pay less attention with lightweight and dynamic convolutions,” in *Proceedings of ICLR*, 2019.
- [19] J. Chen and D. Yang, “Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization,” in *Proceedings of EMNLP*, 2020.
- [20] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” 2019.
- [21] R. Nallapati, B. Zhou, C. dos Santos, C. Gulcehre, and B. Xiang, “Abstractive text summarization using sequence-to-sequence rnns and beyond,” in *Proceedings of SIGNLL*, 2016.
- [22] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “Bertscore: Evaluating text generation with BERT,” in *Proceedings of ICLR*, 2020.
- [23] T. Scialom, P.-A. Dray, S. Lamprier, B. Piwowarski, J. Staiano, A. Wang, and P. Gallinari, “Questeval: Summarization asks for fact-based evaluation,” in *Proceedings of EMNLP*, 2021.
- [24] J. L. Fleiss, “Measuring nominal scale agreement among many raters,” *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.
- [25] J. Kupiec, J. Pedersen, and F. Chen, “A trainable document summarizer,” in *Proceedings of SIGIR*, 1995.
- [26] X. Zhang, F. Wei, and M. Zhou, “Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization,” in *Proceedings of ACL*, 2019.
- [27] Y. Zou, X. Zhang, W. Lu, F. Wei, and M. Zhou, “Pre-training for abstractive document summarization by reinstating source text,” in *Proceedings of EMNLP*, 2020.
- [28] C. Zhu, R. Xu, M. Zeng, and X. Huang, “End-to-end abstractive summarization for meetings,” *arXiv e-prints*, pp. arXiv-2004, 2020.
- [29] Z. Liu, A. Ng, S. L. S. Guang, A. Aw, and N. F. Chen, “Topic-aware pointer-generator networks for summarizing spoken conversations,” *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 814–821, 2019.
- [30] M. Li, L. Zhang, H. Ji, and R. J. Radke, “Keep meeting summaries on topic: Abstractive multi-modal meeting summarization,” in *Proceedings of ACL*, 2019.
- [31] J. Liu, Y. Zou, H. Zhang, H. Chen, Z. Ding, C. Yuan, and X. Wang, “Topic-aware contrastive learning for abstractive dialogue summarization,” in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 1229–1243.
- [32] C.-W. Goo and Y.-N. Chen, “Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts,” in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018.
- [33] Z. Liu, K. Shi, and N. Chen, “Coreference-aware dialogue summarization,” in *Proceedings of SIGDIAL*, 2021.
- [34] M. Li, S. Roller, I. Kulikov, S. Welleck, Y.-L. Boureau, K. Cho, and J. Weston, “Don’t say that! making inconsistent dialogue unlikely with unlikelihood training,” in *Proceedings of ACL*, 2020.
- [35] T. He and J. Glass, “Negative training for neural dialogue response generation,” in *Proceedings of ACL*, 2020, pp. 2044–2058.
- [36] K. Shirai, K. Hashimoto, A. Eriguchi, T. Ninomiya, and S. Mori, “Neural text generation with artificial negative examples,” *arXiv preprint arXiv:2012.14124*, 2020.