Entity-based De-noising Modeling for Controllable Dialogue Summarization

Zhengyuan Liu[†‡], **Nancy F. Chen**[†‡]
[†]Institute for Infocomm Research, A*STAR, Singapore
[‡]CNRS@CREATE, Singapore
{liu_zhengyuan,nfychen}@i2r.a-star.edu.sg

Abstract

Although fine-tuning pre-trained backbones produces fluent and grammatically-correct text in various language generation tasks, factual consistency in abstractive summarization remains challenging. This challenge is especially thorny for dialogue summarization, where neural models often make inaccurate associations between personal named entities and their respective actions. To tackle this type of hallucination, we present an entity-based de-noising model via text perturbation on reference summaries. We then apply this proposed approach in beam search validation, conditional training augmentation, and inference post-editing. Experimental results on the SAMSum corpus show that state-of-the-art models equipped with our proposed method achieve generation quality improvement in both automatic evaluation and human assessment.

1 Introduction

Abstractive dialogue summarization is an emerging research area (Goo and Chen, 2018; Chen et al., 2021). While the data size of available corpora is smaller than that for monological summarization (Carletta et al., 2005; Gliwa et al., 2019), neural approaches have shown promising potential to generate fluent outputs via fine-tuning large-scale contextualized language backbones (Chen and Yang, 2020; Feng et al., 2021). In most corpus constructed for text summarization, only one reference summary is annotated, and models trained via supervised learning on such corpora provide summaries in a general-purpose manner. However, in practice, the generic text summarizers cannot meet the requirements of certain applications and use cases (Fan et al., 2018; Goodwin et al., 2020). For instance, when generating minutes for meeting transcripts, users have their preferences on different personal perspectives. In this case, controllable summarization provides a flexible solution (He et al., 2020) since it allows users to obtain diverse generations. As the aim of dialogue summaries often focuses on *"who did what"* and their narrative flow usually starts with a subject (often persons), the generation process can be modulated by personal named entity planning or prompts (Liu and Chen, 2021). For example, as shown in Figure 1, a controllable system can produce different summaries based on the specific perspective prompts.[1]

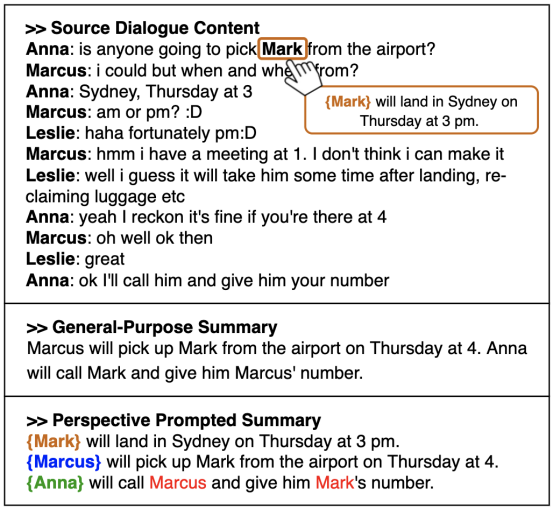

Figure 1: Dialogue summarization examples generated with a general purpose and perspective prompts (labeled in bracket). Note that controllable summaries start with the specified personal named entity's perspective.

However, neural abstractive models often suffer from hallucinations, which lower the reliability of automatic summarization (Zhao et al., 2020; Zhang et al., 2020). In dialogue summarization, this issue commonly involves misaligned personal named entity associations (Lee et al., 2021; Liu et al., 2021b). For instance, as shown in Figure 1, the model upon the prompt *'Anna'* generates the description *"Anna will call Marcus and give him Mark's number"*. While this sentence achieves a high score in word-

[1]Here we use 'prompt' (namely a text conditional signal) under conditional language generation, which is distinct from the task anchor formulated in few-shot/zero-shot 'prompt-based learning' (Liu et al., 2021a).

overlapping metrics such as ROUGE (Lin, 2004), the semantic meaning it conveys is incorrect (according to the conversation, the personal named entities *'Mark'* and *'Marcus'* (colored in red) are misassigned). Such factual inconsistency, the inability to adhere to facts from the source, is a prevalent and unsolved problem (Kryscinski et al., 2019). This limitation is more substantial in controllable scenarios, as models are *required* to condense and paraphrase important contextual information from various personal perspectives.

In this work, we focus on improving the accuracy of personal named entity assignment. Given a source dialogue content, detecting and correcting the errors in a generated summary is similar to the de-noising process adopted in sequence-to-sequence language modeling schemes (Lewis et al., 2020). Therefore, we build an entity-based de-noising model for dialogue summarization via reference summary perturbation and recovery. We then leverage this de-noising model to improve controllable dialogue summarization: (1) At the training stage, we use the de-noising model as a discriminator, to validate beam search candidates under different prompts, and generate factually consistent summaries. Then the validated summaries are added to the training set, which serves as conditional training augmentation. (2) At the inference stage, we use the de-noising model as a corrector, to amend the generated summaries via post-editing. This approach can also be applied to other generic and controllable dialogue summarizers. Experiments are conducted on SAMSum (Gliwa et al., 2019), which consists of multi-turn dialogues and human-written summaries. Empirical results show that our proposed method reduces personal named entity misassignment and achieves improved generation quality on both automatic measures and human evaluation.

2 Related Work

Text summarization is studied in extractive and abstractive paradigms (Gehrmann et al., 2018). In extractive studies, non-neural approaches utilize various linguistic and statistical features via lexical (Kupiec et al., 1995) and graph-based modeling (Erkan and Radev, 2004), and neural approaches bring about substantial improvements via feature-rich distributional representation and hierarchical context modeling (Nallapati et al., 2017; Kedzie et al., 2018). In contrast, abstractive approaches are expected to generate more concise and fluent summaries, which brings about different technical challenges. To foster end-to-end data-driven methods, corpora in news domain (e.g., CNN/Daily Mail (Hermann et al., 2015), NYT (Sandhaus, 2008)) are constructed, and sophisticated neural architectures for abstractive summarization are proposed, such as LSTM-based encoding-decoding (Rush et al., 2015), pointer-generator networks (See et al., 2017), hybrid extractive-abstractive summarizer Gehrmann et al. (2018), and fine-tuning large-scale pre-trained language models (Liu and Lapata, 2019; Lewis et al., 2020). Recently, datasets for summarizing conversations are constructed from meetings (Zhong et al., 2021) or daily chats (Gliwa et al., 2019). Based on the linguistic features of human conversations, many studies pay attention to utilizing conversational analysis for dialogue summarization, such as leveraging dialogue acts (Goo and Chen, 2018), multi-modal features (Li et al., 2019), topic information (Liu et al., 2019), coreference (Liu et al., 2021b), and fine-grained view segmentation with hierarchical modeling (Chen and Yang, 2020).

Controllable language generation introduces auxiliary signals to obtain diverse or task-specific outputs. Such tasks include text style transfer (Shen et al., 2017) and paraphrasing (Iyyer et al., 2018). There are various conditional signal formats, such as categorical labels (Hu et al., 2017), latent representations, semantic or syntactic exemplars (Gupta et al., 2020), and keyword planning (Hua and Wang, 2020). For controllable text summarization, He et al. (2020) and Dou et al. (2021) proposed two generic frameworks in news domain with length constraint and question/entity indicators, and Liu and Chen (2021) proposed personal named entity planning by leveraging the common narrative flow of dialogue summarization.

Tackling hallucinations in abstractive summarization is an essential research topic in making such summaries applicable to real-world scenarios (Kryscinski et al., 2019; Zhao et al., 2020). Reinforcement approaches proposed using factual consistency as optimization reward (Zhang et al., 2020) and post-editing approaches (Kryscinski et al., 2020) focus on correcting summary of general news corpora or facts extracted from an external knowledge base (Iso et al., 2020). For dialogue summarization, Liu and Chen (2021) proposed a binary classifier to detect personal named entity in-

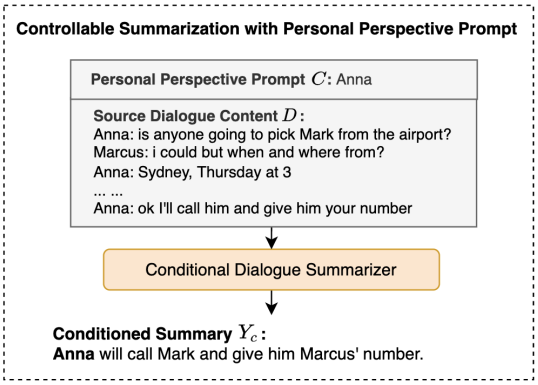

Figure 2: Overview of controllable summarization process. One specific personal named entity is fed to the summarizer as conditional signal.

consistency. Recently, Lee et al. (2021) proposed a post-correction model that can discriminate which type of speaker inconsistency, and revise the output accordingly. In this work, to the best of our knowledge, we are the first to exploit an entity-based de-noising model for abstractive dialogue summarization in both training and inference stages.

3 Controllable Dialogue Summarization

3.1 Task Definition

Here we assume that the input consists of two entries in the controllable setting: a source dialogue D, and a prompt C. The output is the summary text Y, which is a condensed version of the source content D, and starts with the prompt C. Unlike the general-purpose summarization task (Hermann et al., 2015; Gliwa et al., 2019), given one instance of D, the summary Y can be manifested as various outputs conditioned on different choices of C, and are expected to be fluent and factually correct.

3.2 Conditional Entity-based Prompt

In previous studies on controllable document summarization, conditional signals in the form of keywords or descriptive prompts are investigated, and extracted from the source document (He et al., 2020). To summarize multi-turn dialogues, personal named entities that occur in the conversation can be used to form the prompt C for conditional generation (Liu and Chen, 2021). For instance, when writing meeting minutes, with a controllable system, users can obtain diverse generations by choosing different personal named entities, as shown in Figure 1.

In this work, we use the **single entity prompt** for controllable dialogue summarization, as shown in Figure 2. In our reference summary analysis of the SAMSum corpus (Gliwa et al., 2019), the average number of personal named entities (e.g., speaker roles, mentioned persons) in a source dialogue is 2.89. Among these dialogues, 90% human-written summaries start with a personal named entity. In particular, we observed that there is a positional correlation between entity prompts in reference summaries and their occurrence in the source content. As shown in Figure 3, 62% reference summaries start with the first occurred personal named entities in the conversation. This number reaches 94% when we count the first two personal named entities. Therefore, the general-purpose summarizer will follow the same narrative style (namely start with the first speaker or mentioned person), which shares a similar parallel with the position-bias phenomenon studied in news summarization (Kryscinski et al., 2019). Moreover, this positional distribution demonstrates the limited annotation diversity if we only use the reference summary for conditional training.

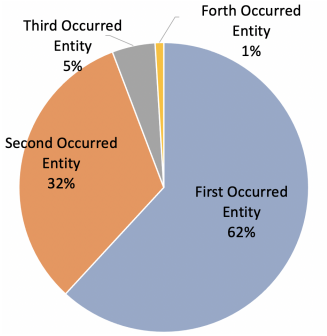

Figure 3: Positional distribution of the personal named entity prompt of reference summaries and their occurrence in the source content.

3.3 Controllable Neural Summarizer

A neural sequence-to-sequence network is applied to build the controllable dialogue summarizer. Its base architecture is a Transformer-based encoding-decoding model, since Transformer (Vaswani et al., 2017) is widely adopted in various natural language processing tasks due to its superior generation performance (Devlin et al., 2019; Lewis et al., 2020).

Encoder: The encoder consists of a stack of Transformer layers. Each layer has two sub-components: a multi-head layer with a self-attention mechanism, and a position-wise feed-forward layer (Equation 1). A residual connection is employed between each pair of the two sub-components, followed by layer normalization (Equation 2).

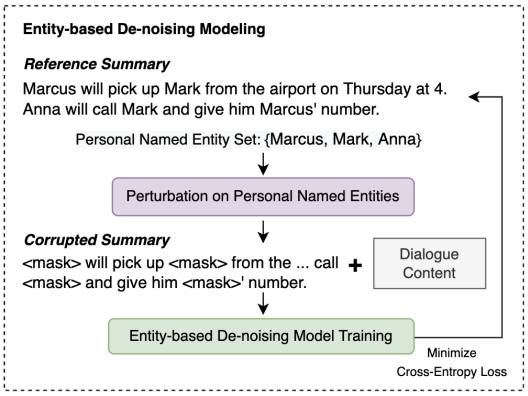

Figure 4: Overview of the entity-based de-noising model. Entity-based text perturbation is conducted on the reference summaries.

$$\widetilde{h}^l = \text{LayerNorm}(h^{l-1} + \text{MHAtt}(h^{l-1})) \quad (1)$$

$$h^l = \text{LayerNorm}(\widetilde{h}^l + \text{FFN}(\widetilde{h}^l)) \quad (2)$$

where l represents the depth of stacked layers, and h^0 is the embedded input sequence. MHAtt, FNN, LayerNorm are multi-head attention, feed-forward and layer normalization components, respectively.

Decoder: The decoder is also a stack of Transformer layers. Aside from the two sub-components in encoding layers, the decoder has another component that performs a multi-head attention over hidden representations from the last encoding layer. Then, the decoder generates tokens from left to right in an auto-regressive manner. Full neural architecture and formula details are described in (Vaswani et al., 2017).

At the training stage, the prompt $C = \{c_0, c_1, .., c_m\}$[2] is concatenated with the source content $D = \{w_0, w_1, ..., w_n\}$ as input, and it is represented as *[<BOS>, C, <EOS>, <BOS>, D, <EOS>]*.[3] To better model utterance boundary representation, we added a special token *'<u>'* as the utterance delimiter in D. [4] The summarizer learns to generate the ground truth $Y = \{y_0, y_1, ..., y_t\}$ by condensing the information of dialogue context conditioned on the prompt. The loss of maximizing the log-likelihood on the ground truth is formulated as:

$$loss(\theta) = -\Sigma_i \log(p(y_i|y_{<i}, D, C; \theta)) \quad (3)$$

[2]In our setting, while the prompt is a single personal named entity, it can be multiple tokens after the subword tokenization.

[3]Tokens of <BOS> and <EOS> defined in *'BART-large'* are <s> and </s> respectively, and can be changed according to other language backbones.

[4]The special token *'<u>'* is added to the vocabulary, and we initialize its token embedding by averaging the embedding vectors of '<s>', comma, and period.

Sample Type	Number
Training Set (14732 Samples)	
Mean/Std. of Dialogue Turns	11.7 (6.45)
Mean/Std. of Dialogue Length	124.5 (94.2)
Mean/Std. of Summary Length	23.44 (12.72)
Validation Set (818 Samples)	
Mean/Std. of Dialogue Turns	10.83 (6.37)
Mean/Std. of Dialogue Length	121.6 (94.6)
Mean/Std. of Summary Length	23.42 (12.71)
Test Set (819 Samples)	
Mean/Std. of Dialogue Turns	11.25 (6.35)
Mean/Std. of Dialogue Length	126.7 (95.7)
Mean/Std. of Summary Length	23.12 (12.20)

Table 1: Data Statistics of the dialogue summarization dataset SAMSum (Gliwa et al., 2019).

where D, C, y, θ denotes the dialogue content, conditional prompt, targeted summary sequence, and the trainable parameter set, respectively. i is decoding time-step, and ranges from 1 to t. During inference, the model creates a summary based on a specific perspective prompt, and is coherent with the context of the input conversation.

4 Entity-based De-noising Modeling

While existing abstractive neural models achieve state-of-the-art performance on quantitative evaluation, factual inconsistency remains a prevalent and unsolved problem (Kryscinski et al., 2019; Zhang et al., 2020). In both document and dialogue summarization, it has been demonstrated that a certain proportion of abstractive summaries contain hallucinated statements (Zhao et al., 2020; Khalifa et al., 2021). Such hallucinations raise concerns about the usefulness and reliability of automatic summarization, and are challenging to eradicate in neural approaches due to the implicit nature of learning representations.

In dialogue summarization, the misassignment of personal named entities significantly affects generation quality (Lee et al., 2021; Liu and Chen, 2021). Inspired by the de-noising sequence-to-sequence pre-training schemes (Lewis et al., 2020; Raffel et al., 2020), here we propose an entity-based de-noising model to detect and recover the incorrect personal named entity tokens. Compared with the binary classifier for factual inconsistency (Liu and Chen, 2021), the sequence-to-sequence framework supports revising the summaries via post-editing.

4.1 De-noising Sample Construction

To construct training samples for entity-based de-noising, we conduct text perturbation on the ref-

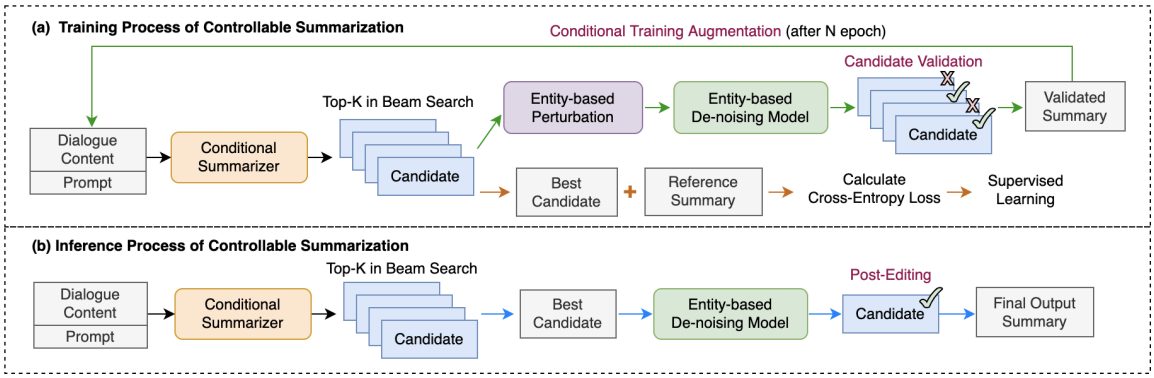

Figure 5: Overview of the controllable summarization framework equipped with entity-based de-noising modeling: (a) Training process with supervised learning (in orange arrow), and with beam search validation and conditional sample augmentation (in green arrow); (b) Inference process with post editing (in blue arrow).

erence summaries. As shown in Figure 4, given a reference summary Y, we obtain a corrupted version $\widetilde{Y}$ via entity masking and substitution. More specifically, we first extract the full list of personal named entities from each source dialogue, and then mask them or replace them with another entity at a random rate (p_{noise}=0.5). Additionally, to reduce the positional imbalance caused by labeling correlation described in Section 3.2, we shuffle the summary sentences at a random rate (p_{shuffle}=0.5) .

4.2 De-noising Model Training

We fine-tuned the sequence-to-sequence language backbone *BART-large* (Lewis et al., 2020) for de-noising modeling. Given a dialogue D and a corrupted summary $\widetilde{Y}$, the input is represented as *[<BOS>, $\widetilde{Y}$, <EOS>, <BOS>, D, <EOS>]*.[5] As shown in Figure 4, the model is applied to generate the reference summary Y, and is optimized by minimizing the cross-entropy loss. Since the text perturbation is conducted especially on personal named entities, it encourages the de-noising backbone to model features such as *"who-did-what"* and speaker interactions. Moreover, unlike the left-to-right auto-regressive summary generation, the de-noising backbone can utilize the bi-directional context of both dialogue and summary sequence, and it achieves a 0.92 sample-level accuracy on the validation set, which is a reasonable performance for follow-up steps.

5 Leveraging De-noising Modeling

In this section, we then elaborate on how to leverage the entity-based de-noising model for controllable dialogue summarization.

5.1 Beam Search Candidate Validation

The de-noising model can be used as a reference-free discriminator to validate the beam search candidates. Following previous work on two-stage summary ranking (Liu and Liu, 2021), we use diverse beam search (Vijayakumar et al., 2016) as the sampling strategy. As shown in Figure 5 (a) and Figure 6, for each candidate generated in beam search, we mask all the personal named entities, and feed it to the de-noising model. If the recovered output is identical to the unaltered candidate, it is regarded as a validated summary without any personal named entity misassignment. Moreover, given a pair of names concatenated with '*and*', we consider their permutations are the same (e.g, '*Tom and John*', '*John and Tom*').

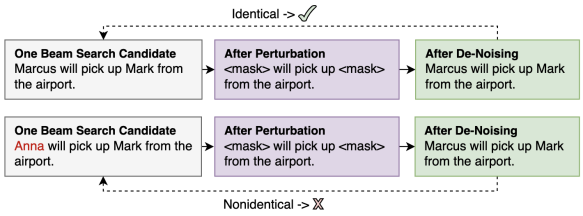

Figure 6: One example of beam search candidate validation. Two beam search candidates are validated by the de-noising discriminator, and one misassignment is detected ('Anna' in red).

5.2 Conditional Training Augmentation

One major challenge of training models for controllable dialogue summarization is the lack of diverse annotation, as each source content only has one reference summary in existing corpora. Moreover, due to the positional correlation of entity prompts

[5]Tokens of <BOS> and <EOS> defined in *'BART-large'* are <s> and </s> respectively, and can be changed according to other language backbones.

in human-written summaries (Section 3.2), their corresponding conditional samples will present an imbalanced prompt distribution, and cause unnecessary inductive bias in data-driven approaches.

In this work, we address this issue by introducing weak self-supervision (Karamanolakis et al., 2021), and use the summarizer's intermediate generation as additional training samples. In other controllable language generation studies like text style transfer, self-supervised sample selection adopts metrics such as sentiment polarity score (Luo et al., 2019); here we use the entity-based consistency. As shown in Figure 5 (a), after N training epoch, we re-run the model on the original training set, obtain summaries upon perspective prompts which are distinct from that of the reference, and validate them by the de-noising model (as in Section 5.1), then the validated samples which rank highest in beam search are used as additional training data. In our experiments on SAMSum, we conducted the augmentation from the third epoch (when the summarizer produces reasonable results with automatic metrics), and 30% of the training set contribute a conditional augmented sample.

5.3 Inference with Post-Editing

In addition, since the de-noising model learns to correct the entity-based perturbation, it can also be used for summary post-editing, which is an effective method to improve the generation quality commonly applied in machine translation (Popović and Arčan, 2016). As shown in Figure 5 (b), at the inference stage, the best candidate selected from beam search is fed to the de-noising model, then we obtain the final summary where the misassigned entities are corrected. It is noteworthy to mention that, since the post-editing here focuses on personal named entity correction, it is not straightforward to observe the performance improvement via automatic evaluation metrics such as ROUGE, and we thus conduct a human evaluation. Moreover, the post-editing is a general process to extend to other dialogue summarization systems (see experimental results in Section 6.6).

6 Experiments and Results

6.1 Experimental Corpus

Experiments are conducted on SAMSum (Gliwa et al., 2019), which contains multi-turn daily conversations with human-written summaries in a general-purpose manner. Details of the dataset are shown in Table 1. We retain the original text content of conversations such as cased words, emoticons, and special tokens, and pre-process them using sub-word tokenization (Lewis et al., 2020). Since the positional embedding of our Transformer-based model can support 1,024 input length, none of the samples are truncated.

6.2 De-noising Model Configuration

The *'BART-large'* (Lewis et al., 2020) is used to build the entity-based de-noising model. The number of encoder layers, decoder layers, input, and hidden dimension and $12/12/1024$, respectively. The learning rate was set at $2e{-}5$. *AdamW* optimizer (Loshchilov and Hutter, 2019) was used with weight decay of $1e{-}3$ and a linear scheduler. Drop-out (Srivastava et al., 2014) ($rate$=0.1) was set as in the original *BART* configuration. Text perturbation described in Section 4.1 is conducted on the SAMSum dataset for training and validation.

6.3 Summarization Model Configuration

For controllable dialogue summarization, the language backbone *BART* (Lewis et al., 2020) is applied. The number of encoder layers, decoder layers, input and hidden dimension are $6/6/768$ for the *'BART-base'*, and $12/12/1024$ for the *'BART-large'* and *'CTRLsum'*. *AdamW* optimizer (Loshchilov and Hutter, 2019) was used with learning rate of $3e{-}5$, weight decay of $1e{-}3$, and a linear learning rate scheduler. Drop-out (Srivastava et al., 2014) rate was set at 0.1. Diverse beam search (Vijayakumar et al., 2016) is adopted with group number 5 and beam size 10. For augmentation samples, we added a weighted loss (λ=0.15).

The trainable parameter size is 139M of the *'BART-base'*, and 406M of the *'BART-large'*. Batch size and epoch number were set at 8. Best checkpoints were selected based on validation results of ROUGE-2 F1 score. All models were implemented with PyTorch (Paszke et al., 2019) and HuggingFace Transformers[6]. All experiments were running on a single Tesla A100 GPU with 40G memory.

6.4 Evaluation Metrics

Extensive metrics are used for quantitative evaluation: (1) We adopt **ROUGE-1**, **ROUGE-2**, and **ROUGE-L** (Lin, 2004), which are customary in summarization tasks via counting lexical overlap, and `Py-rouge` package is employed following

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

	ROUGE-1			ROUGE-2			ROUGE-L			SimCSE	B-Score	EACC
	F	P	R	F	P	R	F	P	R			
CTRLsum BART$_{large}$ (CNN/DM)	33.8	43.5	32.6	10.5	14.2	10.1	33.6	41.1	32.1	61.8	-7.19	71.2
CTRLsum BART$_{large}$ (SAMSum)	53.8	55.9	57.6	29.9	31.2	31.9	52.4	53.6	55.1	79.2	-4.79	84.2
+ Beam Search Valid-Augment	54.2	57.7	56.4	30.3	32.9	31.3	52.9	55.4	54.3	79.4	-4.71	88.7
Conditional BART-base	51.3	57.1	51.8	27.2	30.4	27.5	50.4	54.6	50.3	76.2	-5.36	74.4
+ Beam Search Valid-Augment	51.5	58.4	50.7	27.6	31.4	27.3	50.7	56.1	49.7	76.5	-5.20	79.8
Conditional BART-large	53.8	61.6	52.6	30.2	35.1	29.4	52.9	59.1	51.5	78.4	-5.23	84.7
+ Beam Search Valid-Augment	54.2	58.8	55.2	30.4	33.4	30.8	53.0	56.5	53.5	78.9	-4.98	86.2

Table 2: **Results on reference prompt generation** (matching training and test condition). *F*, *P*, *R* are F1 measure, precision, and recall. *B-Score* and *EACC* denotes BARTScore (Yuan et al., 2021) and entity-based accuracy. *CTRLsum* is a generic controllable summarizer for news (He et al., 2020), and we further fine-tuned it on SAMsum.

	ROUGE-1			ROUGE-2			ROUGE-L			SimCSE	B-Score	EACC
	F	P	R	F	P	R	F	P	R			
CTRLsum BART$_{large}$ (CNN/DM)	34.8	50.1	32.7	10.7	17.7	9.7	31.7	42.8	29.2	60.0	-7.99	71.7
CTRLsum BART$_{large}$ (SAMSum)	54.3	56.9	57.8	28.0	29.3	29.9	50.5	52.1	52.9	78.8	-4.85	67.8
+ Beam Search Valid-Augment	55.1	58.2	56.5	28.7	30.7	29.4	50.7	52.8	51.7	79.0	-4.82	77.3
Conditional BART-base	51.5	58.2	50.1	29.4	29.7	25.4	50.9	56.4	48.5	75.0	-5.44	64.1
+ Beam Search Valid-Augment	52.6	59.4	50.7	28.0	31.5	27.1	50.4	56.3	49.1	75.8	-5.23	72.2
Conditional BART-large	55.1	62.3	53.3	29.5	33.5	29.3	53.2	59.2	52.3	78.2	-5.04	68.3
+ Beam Search Valid-Augment	55.7	61.5	56.2	30.2	32.5	30.5	53.6	57.1	53.7	79.5	-4.91	77.2

Table 3: **Results on distinct prompt generation** (simulating a practical use case). *F*, *P*, *R* are F1 measure, precision, and recall. *B-Score* and *EACC* denotes BARTScore (Yuan et al., 2021) and entity-based accuracy.

	ROUGE-1			ROUGE-2			ROUGE-L			SimCSE	B-Score	EACC
	F	P	R	F	P	R	F	P	R			
CTRLsum BART$_{large}$ (CNN/DM)	32.4	42.7	30.7	9.5	13.3	9.0	32.0	39.9	30.1	59.9	-7.57	77.9
CTRLsum BART$_{large}$ (SAMSum)	52.2	53.2	57.1	27.0	27.8	29.5	48.7	49.2	52.2	77.7	-4.79	84.4
+ Beam Search Valid-Augment	52.2	54.0	56.6	27.3	28.7	29.3	48.6	49.7	51.7	77.2	-4.91	87.4
General-Purpose BART-base	50.5	54.8	51.9	25.2	27.4	26.1	47.7	50.7	48.5	75.1	-5.25	78.0
Conditional BART-base	50.0	55.2	50.5	24.8	27.5	25.1	47.1	50.8	47.2	74.0	-5.35	72.9
+ Beam Search Valid-Augment	49.4	55.6	49.1	24.7	28.1	24.7	46.9	51.5	46.4	73.4	-5.45	78.2
General-Purpose BART-large	53.0	57.2	54.5	28.1	30.8	28.7	49.8	53.1	50.5	77.6	-5.11	88.5
Conditional BART-large	52.6	58.1	53.3	27.7	30.9	27.6	49.1	53.1	49.0	76.3	-5.28	85.3
+ Beam Search Valid-Augment	51.9	54.7	54.4	27.4	29.3	28.6	48.1	50.1	49.7	75.9	-5.05	88.1

Table 4: **Results on generation without prompt** (simulating the non-conditional scenario). *F*, *P*, *R* are F1 measure, precision, and recall. *B-Score* and *EACC* are BARTScore (Yuan et al., 2021) and entity-based accuracy.

(Gliwa et al., 2019; Feng et al., 2021). (2) We use the **SimCSE** (Gao et al., 2021) and **BARTScore** (Yuan et al., 2021) as semantic-oriented metrics via counting embedding similarity, which have empirically shown to match the human perception. (3) We calculate the **Entity-based Accuracy** by applying a separate de-noising model on final outputs, which is formulated as the proportion of samples that all personal named entities are correctly generated.

6.5 Quantitative Evaluation Results

6.5.1 Reference Prompt Generation

Firstly, to evaluate the controllable generation under the matched training and test condition, we obtain summaries with the same entity prompt as reference summaries, and this matched condition serves as the benchmarked performance for conditional learning (Liu and Chen, 2021). As shown in Table 2, adding the conditional data augmentation to *'BART-base'* and *'BART-large'* brings improvement on both lexical and semantic metrics, which significantly improves entity-based accuracy. We also evaluate a generic controllable model *'CTRLsum'* (He et al., 2020) that provides state-of-the-art results on the news domain. We observed that *'CTRLsum'* benefits from domain adaptation from news to dialogue (on all fronts) when running experiments on the SAMSum corpus, and we obtained further improvements when adopting the proposed training augmentation.

Source Dialogue Content: Pamela: Have you tried applying for the RGS grant? <u> Zoe: I did last year. <u> Zoe: Wasn't lucky... <u> Xiara: Me neither <u> Pamela: Do you know if I have to be a member? <u> Zoe: It depends on the award <u> Pamela: Ok I'll have a look at their website.
Reference Summary: **Pamela** will look at their website to apply for the RGS grant. Zoe and Xiara applied last year but they weren't lucky. Sometimes you have to be a member to apply. **Generation with General-Purpose (BART-large):** **Zoe** applied for the RGS grant last year, but didn't get lucky. Pamela will look at their website. **Generation with Prompt (BART-large):** **Pamela**, **Zoe** and **Xiara** didn't apply for the RGS grant last year. **Generation with Prompt (BART-large + Valid-Augment):** **Pamela** will have a look at the RGS website to apply for the grant.
Source Dialogue Content: Ivan: hey eric <u> Eric: yeah man <u> Ivan: so youre coming to the wedding <u> Eric: your brother's <u> Ivan: yea <u> Eric: i dont know mannn <u> Ivan: YOU DONT KNOW?? <u> Eric: i just have a lot to do at home, plus i dont know if my parents would let me <u> Ivan: ill take care of your parents <u> Eric: youre telling me you have the guts to talk to them XD <u> Ivan: thats my problem <u> Eric: okay man, if you say so <u> Ivan: yea just be there <u> Eric: alright.
Reference Summary: **Eric** doesn't know if his parents let him go to Ivan's brother's wedding. Ivan will talk to them. **Generation with General-Purpose (BART-large):** **Ivan** is going to **Eric**'s brother's wedding. Eric doesn't know if he can come because he has a lot to do at home. Ivan will talk to his parents. **Generation with General-Purpose (BART-large + Post-Editing):** **Eric** is going to Ivan's brother's wedding. Eric doesn't know if he can come because he has a lot to do at home. Ivan will talk to his parents.

Table 5: Two examples of dialogues in SAMSum, and summaries generated by different models. <u> is the utterance delimiter. Text in blue denotes beginning or prompt entities. Text in red denotes the factual inconsistency.

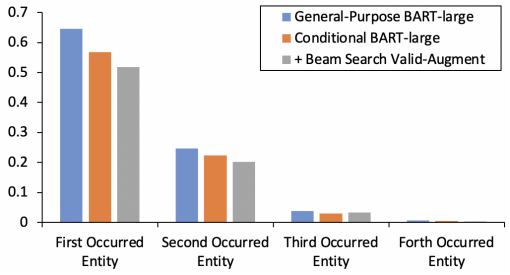

Figure 7: Positional distribution of start entity generated by different models without prompt.

6.5.2 Distinct Prompt Generation

As the single reference summary cannot be readily used for diverse conditional evaluation, to simulate the practical controllable generation scenario, we build a sub-set with distinct prompts (119 of 819 test samples), where the generation by a general-purpose *'BART-large'* and reference summaries start with different personal named entities. As shown in Table 3, all models (*'BART-base'*, *'BART-large'*, and *'CTRLsum'*) with the proposed method achieve higher performance on all fronts, and their entity-based consistency has a relative 12% gain.

6.5.3 Generation without Prompt

While controllable summarizers require a prompt as part of the input, we also obtained summaries without any entity indicator to simulate the general-purpose summarization scenario. As shown in Table 4, models trained in a conditional manner achieve comparable but slightly lower scores. As shown in Figure 7, we speculate that this is because summaries generated by conditionally-trained models present a more balanced entity distribution.

Model	Error Rate
Conditional BART-large	0.37
+ Beam Search Valid-Augment	0.27
+ Inference Post-Editing	0.23

Table 6: Human assessment on entity-based factual consistency of distinct prompt generation.

6.5.4 Results after Inference Post-Editing

For all three generation types shown in Table 2, Table 3, and Table 4), we observed that adopting inference post-editing does not affect the lexical and semantic scores (as it only changes a few tokens), but this post-processing step can improve entity-based consistency by 7% relatively.

Moreover, following previous work (Liu and Chen, 2021), we incorporated dialogue coreference information for the controllable generation, and it is effective to improve the generation quality such as entity accuracy (see results in Appendix).

6.6 Human Assessment on Entity-based Factual Consistency

We further conducted two qualitative evaluations via human assessment. At each time, 30 samples are randomly chosen from the test set and their corresponding summaries from different summarizers. Participants are asked to read the dialogue and summaries, and judge if any personal named entity is misassigned.

For controllable summarization, we evaluate the outputs upon the **distinct prompt generation** (de-

Model	Error Rate
General-Purpose BART-large	0.33
+ Inference Post-Editing	0.26

Table 7: Human assessment on entity-based factual consistency of general-purpose models.

scribed in Section 6.5.2). As shown in Table 6, we observe that the sample-level error rate drops from 0.37 to 0.27 (22% relatively) with the conditional training augmentation, and this is consistent with automatic entity-based accuracy results (see examples in Table 5), and it further drops to 0.23 after the post-editing.

Since the inference post-editing described in Section 5.3 can also be adopted on general-purpose summarizers, we conduct a human assessment on the **non-conditional generation**: we fine-tune a *'BART-large'* which serves as the state-of-the-art baseline on the original SAMSum corpus, and feed its generation to the entity-based de-noising model for post-editing. As shown in Table 7, we observe that the sample-level error rate drops from 0.33 to 0.26 (25% relatively) with the post-editing (see examples in Table 5).

7 Conclusion

In this paper, we focused on reducing incorrect assignments of personal named entities in dialogue summarization. We proposed an entity-based de-noising model, and applied it to beam search validation, conditional training augmentation, and inference post-editing (which can be used for non-conditional and conditional summarization). Experimental results demonstrated that our proposed method improves performance in both lexical and semantic evaluation metrics and is beneficial to entity-based factual consistency in both automatic and human evaluations. Future work can be extending it to pronoun tokens and other entity types.

Acknowledgments

This research was supported by funding from the Institute for Infocomm Research (I2R) under A*STAR ARES, Singapore, and by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. We thank Ai Ti Aw for the insightful discussions. We also thank the anonymous reviewers for their precious feedback to help improve and extend this piece of work.

References

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. 2005. The ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer.

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of EMNLP 2020*, pages 4106–4118. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL 2019*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. Gsum: A general framework for guided neural abstractive summarization. In *Proceedings of NAACL 2021*, pages 4830–4842. Association for Computational Linguistics.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Angela Fan, David Grangier, and Michael Auli. 2018. Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54. Association for Computational Linguistics.

Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. 2021. Language model as an annotator: Exploring DialoGPT for dialogue summarization. In *Proceedings of ACL 2021*, pages 1479–1491. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of EMNLP 2021*, pages 6894–6910.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of EMNLP 2018*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Chih-Wen Goo and Yun-Nung Chen. 2018. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 735–742. IEEE.

Travis Goodwin, Max Savery, and Dina Demner-Fushman. 2020. Towards zero shot conditional summarization with adaptive multi-task fine-tuning. In *Proceedings of EMNLP 2020*, pages 3215–3226.

Prakhar Gupta, Jeffrey P Bigham, Yulia Tsvetkov, and Amy Pavel. 2020. Controlling dialogue generation with semantic exemplars. *arXiv preprint arXiv:2008.09075*.

Junxian He, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2020. Ctrl-sum: Towards generic controllable text summarization. *arXiv preprint arXiv:2012.04281*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of NeurIPS 2015*, pages 1693–1701.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *Proceedings of ICML 2017*, pages 1587–1596.

Xinyu Hua and Lu Wang. 2020. PAIR: Planning and iterative refinement in pre-trained transformers for long text generation. In *Proceedings of EMNLP 2020*, pages 781–793. Association for Computational Linguistics.

Hayate Iso, Chao Qiao, and Hang Li. 2020. Fact-based Text Editing. In *Proceedings of ACL 2020*, pages 171–182, Online. Association for Computational Linguistics.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of NAACL 2018*, pages 1875–1885.

Giannis Karamanolakis, Subhabrata Mukherjee, Guoqing Zheng, and Ahmed Hassan Awadallah. 2021. Self-training with weak supervision. In *Proceedings of NAACL 2021*, pages 845–863, Online. Association for Computational Linguistics.

Chris Kedzie, Kathleen McKeown, and Hal Daume III. 2018. Content selection in deep learning models of summarization. In *Proceedings of EMNLP 2018*, pages 1818–1828, Brussels, Belgium. Association for Computational Linguistics.

Muhammad Khalifa, Miguel Ballesteros, and Kathleen McKeown. 2021. A bag of tricks for dialogue summarization. In *Proceedings of EMNLP 2021*, pages 8014–8022, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the EMNLP 2019*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of EMNLP 2020*, pages 9332–9346.

Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '95, page 68–73, New York, NY, USA. Association for Computing Machinery.

Dongyub Lee, Jungwoo Lim, Taesun Whang, Chanhee Lee, Seungwoo Cho, Mingun Park, and Heuiseok Lim. 2021. Capturing speaker incorrectness: Speaker-focused post-correction for abstractive dialogue summarization. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 65–73, Online and in Dominican Republic. Association for Computational Linguistics.

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 ACL 2020*, pages 7871–7880. Association for Computational Linguistics.

Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of ACL 2019*, pages 2190–2196, Florence, Italy. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of EMNLP 2019*, pages 3721–3731, Hong Kong, China. Association for Computational Linguistics.

Yixin Liu and Pengfei Liu. 2021. Simcls: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of ACL-IJCNLP 2021*, pages 1065–1072.

Zhengyuan Liu and Nancy Chen. 2021. Controllable neural dialogue summarization with personal named entity planning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 92–106, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhengyuan Liu, Angela Ng, Sheldon Lee, Ai Ti Aw, and Nancy F Chen. 2019. Topic-aware pointer-generator networks for summarizing spoken conversations. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 814–821. IEEE.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021b. Coreference-aware dialogue summarization. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 509–519, Singapore and Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *The International Conference on Learning Representations (ICLR 2019)*.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Xu Sun, and Zhifang Sui. 2019. A dual reinforcement learning framework for unsupervised text style transfer. In *IJCAI*.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of AAAI 2017*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of NeurIPS 2019*, pages 8026–8037.

Maja Popović and Mihael Arčan. 2016. PE2rr corpus: Manual error annotation of automatically pre-annotated MT post-edits. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 27–32, Portorož, Slovenia. ELRA.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of EMNLP 2015*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

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*, pages 1073–1083.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Proceedings of NeurIPS 2017*, pages 6830–6841.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NeurIPS 2017*, pages 5998–6008.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. 2020. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of ACL 2020*, pages 5108–5120. Association for Computational Linguistics.

Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2237–2249. Association for Computational Linguistics.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the NAACL*, pages 5905–5921, Online. Association for Computational Linguistics.

A Appendix

	ROUGE-1			ROUGE-2			ROUGE-L			SimCSE	B-Score	EACC
	F	P	R	F	P	R	F	P	R			
Conditional BART-large	53.8	61.6	52.6	30.2	35.1	29.4	52.9	59.1	51.5	78.4	-5.23	84.7
+ Coreference Information	54.3	57.9	56.8	30.2	32.6	31.5	52.6	55.3	54.3	79.5	-4.72	85.8
+ Beam Search Valid-Augment	54.1	56.0	58.4	30.5	32.1	32.4	52.3	53.5	55.3	79.2	-4.69	86.8
+ Inference Post-Editing	54.2	56.6	57.7	30.3	32.1	32.3	52.5	54.2	55.1	80.2	-4.59	98.2

Table 8: **Additional experimental results on reference prompt generation** (matching training and test condition). We incorporated dialogue coreference information following previous work (Liu and Chen, 2021). *F*, *P*, *R* are F1 measure, precision, and recall. *B-Score* and *EACC* denotes BARTScore and entity-based accuracy.

	ROUGE-1			ROUGE-2			ROUGE-L			SimCSE	B-Score	EACC
	F	P	R	F	P	R	F	P	R			
Conditional BART-large	55.1	62.3	53.3	29.5	33.5	29.3	53.2	59.2	52.3	78.2	-5.04	68.3
+ Coreference Information	54.9	58.5	57.2	29.9	32.1	31.0	51.5	53.6	53.1	79.2	-4.90	76.1
+ Beam Search Valid-Augment	55.5	55.8	59.5	29.1	29.5	31.1	51.7	51.8	54.5	78.5	-4.58	77.7
+ Inference Post-Editing	55.1	55.3	59.1	27.6	27.7	29.8	50.0	49.4	52.5	78.5	-4.57	97.9

Table 9: **Additional experimental results on distinct prompt generation** (simulating a practical use case). We incorporated dialogue coreference information following previous work (Liu and Chen, 2021). *F*, *P*, *R* are F1 measure, precision, and recall. *B-Score* and *EACC* denotes BARTScore, and entity-based accuracy.

	ROUGE-1			ROUGE-2			ROUGE-L			SimCSE	B-Score	EACC
	F	P	R	F	P	R	F	P	R			
Conditional BART-large	52.6	58.1	53.3	27.7	30.9	27.6	49.1	53.1	49.0	76.3	-5.28	85.3
+ Coreference Information	52.2	51.7	58.4	27.1	27.0	30.5	47.9	47.8	52.8	77.3	-4.75	87.2
+ Beam Search Valid-Augment	52.0	51.3	59.1	26.9	26.8	30.6	47.6	46.6	52.8	77.1	-4.69	85.3
+ Inference Post-Editing	51.9	51.2	59.0	26.9	26.9	30.5	47.6	46.7	52.7	76.9	-4.38	98.1

Table 10: **Additional experimental results on generation without prompt** (simulating the non-conditional scenario). We incorporated dialogue coreference information following previous work (Liu and Chen, 2021). *F*, *P*, *R* are F1 measure, precision, and recall. *B-Score* and *EACC* are BARTScore, and entity-based accuracy.