Qi Jia, Yizhu Liu, Siyu Ren and Kenny Q. Zhu
Shanghai Jiao Tong University
800 Dongchuan Road, Shanghai, China, 200240
Jia_qi@sjtu.edu.cn

July 30, 2023

Dr. Albert Zomaya, Editor-in-Chief
ACM Computing Surveys

Dear Dr. Albert Zomaya and Reviewers,

We appreciate the opportunity to revise our manuscript. Thank you for the editors' and reviewers' comments concerning our manuscript entitled "Taxonomy of Abstractive Dialogue Summarization: Scenarios, Approaches and Future Directions" (ACM CSUR-2022-0082.R1). Those comments are all valuable and very helpful for revising and improving our paper. A point-by-point response to the comments is below, and we also highlight the changes to the manuscript using colored text. We believe that the revisions prompted by these comments have strengthened our manuscript.

Reviewer 1

1. The paper has a very good overall coverage of the dialogue summarization task. But one minor issue is that I kind of feel the evaluation metrics section could be improved. Although current version indeed lists some evaluation metrics, but 1. it does not explain why these metrics are suitable for the task 2. now the meta-evaluation of summarization metrics is becoming more and more important, as ROUGE often shows a conflict against human evaluations. It would be good it the survey can also covers this topic.

   AUTHORS' RESPONSE: We improved the section for evaluation metrics by considering these two comments as follows:
   1. We add intuitions of using/designing these evaluation metrics considering the characteristics of task-oriented dialogue summarization(TDS) and open-domain dialogue summarization(ODS) (see Sec. 7.2):

   - Page 26 Line 32: "Instead of comparing only with the whole reference summary, most research for TDS only considers key words/phrases while ignoring other common words for measuring the information coverage. In other words, evaluation for TDS emphasizes the coverage of key information which are generally domain-specific terms and can be easily recognized....."
   - Page 27 Line 1: "ODS pays less attention to information coverage due to the higher subjectivity on salient information selection. Instead, measuring the factual consistency of generations gains increasing attention.....Information correctness of the generated summary is also important for TDS. For instance, ..."

2. We agree that the meta-evaluation of summarization metrics is of great importance. However, due to the limited research for dialogue summarization under this topic, we put the discussion of it in Sec. 8.2 Future Directions:

- Page 32 Line 39: "Evaluation metrics are significant, which guide the improvement directions for upcoming models. However, widely used evaluation metrics in Sec. 7.2 are all borrowed from document summarization tasks and their effectiveness is unverified. Recent work from Gao and Wan [54] re-evaluated 18 evaluation metrics and did a unified human evaluation with 14 dialogue summarization models on SAMSum dataset.....automatic metrics specially designed for dialogue summarization are urgently needed."
- Page 33 Line 8: "Factual errors caused by the mismatch between speakers and events are common as a result of complicated discourse relations among utterances in dialogues.....With the strong generation ability of current LLMs, there's also a doubt that whether the previous taxonomy of error types and evaluation metrics is still suitable. In a word, both meta-evaluation benchmarks and evaluation methods call for innovations."

Reviewer 2

1. While the overall survey is quite impressive and comprehensive, I could not find relevant categorization of Long Dialog summarization techniques which is an emerging field. Authors could provide additional information regarding the techniques, and evaluation strategies for long dialogs, which do not fit the standard context length of many state of the art models.

   AUTHORS' RESPONSE: In this survey, we don't classify the dialogue summarization tasks according to the input dialogue length since both short and long dialogues share a number of features. We do contain approaches, evaluation benchmarks, and discussions for long dialogue summarization, which helps to solve the problem of the sequence length limitation of many state-of-the-art models. More details are as follows:

   - Section 3 Overview of approaches, Page 7 Line 20: "Besides such flat and sequential modeling, hierarchical modeling is another representative design as shown in Figure 3, which is usually favored by longer dialogues. "
   - Section 4.2 Inter-utterance features, Page 12 Line 18: "Partitions ..... reduce the requirements on GPU memory with shorter input lengths, which are especially preferred for long dialogue summarization."
   - Section 4.4 Summary and Opinions, Page 16 Line 3: "Manipulating the input and output by adding annotations or data reformulation ..... For the hierarchical or more complicated structures, researchers tend to reformulate the dialogue into different segments, especially for long dialogues [10, 148, 209] or reordering utterances considering graph features."
   - Section 7.1 Datasets: We contain the average dialogue length(DW) and summary length(SW) in Table 1 and Table 2. Besides, we have: (Page 26 Line 22) "Datasets with more than 2,048 dialogue words, which is the upper bound of the input length of most pre-trained language models, are suitable for research on long dialogue summarization. They contain both open-domain datasets and task-oriented datasets."
   - Section 8.1 Paper analysis, Page 30 Line 10: "Partitions are extremely effective for TDS where dialogues are usually long with inherent semantic transitions, such as agendas for meetings and

domain shifts in customer service ....."

- Section 8.2 Future directions, Page 31 Line 3: "Multi-session dialogue summarization is required when conversations occur multiple times among the same group of speakers ..... However, current approaches generally break down the long dialogue and summary into shorter chunks due to the limitation of current models. For task-oriented scenarios, it is also common in real life. For example, the customer may repeatedly ask for help from the agent for the same issue that hasn't been solved before. An updated summary considering all of the questions and the latest answers can remind the long dialogue history of both participants and facilitate the negotiation process. "

Reviewer 3

1. Insights expected from survey papers: Was there a response to this stated weakness in the previous review? I might have missed it if the authors already addressed this aspect. "While there is a thorough compilation of work in abstractive dialogue summarization, this paper does not provide insights that a high quality survey paper is expected to provide."

   AUTHORS' RESPONSE: We have a subsection named "Summary and Opinions" at the end of Sec. 4, Sec. 5 and Sec. 6 respectively, containing a summary of the corresponding class of approaches and an analysis of their advantages and short-comings. Moreover, in Sec 8, "Analysis and Future Directions", we provide a statistical analysis of papers together with observations and insights, which also leads to possible future directions in the later sub-section. Representative details are as follows:

   - Sec. 4.4, Page 16 Line 1: "The features mentioned above are summarized in Figure 5. They are mainly injected into vanilla models in three ways: Manipulating the input and output..... Modifying the model architecture or hidden states..... Adding additional training targets.....The advantages and disadvantages of injecting pre-processed features are as follows:.....Features collected by labelers on other dialogue understanding tasks capture the essence of these tasks and also establish connections with various aspects of dialogue analysis. Therefore, leveraging such features is a good way to alleviate the human labeling burden.....Error propagation exists in these dialogue summarization approaches. Incorrect features hinder the understanding of dialogues and lead to poor summaries."
   - Sec. 5.4, Page 19 Line 15: "The tasks mentioned above are in Figure 6. Most of these self-supervised tasks are adopted in two ways: Cooperating with the vanilla generation task under different training paradigms.....Training an isolated model for different purposes.....The advantages and disadvantages of designing self-supervised tasks are as follows: Most self-supervised tasks take advantage of self-supervision to train the model. They don't need to go through the expensive and time-consuming annotation process for collecting high-quality labels, and avoid the domain transfer problems of transferring labelers trained on the labeled domain to the target summarization domain..... Different self-training tasks are not always compatible and controllable. It is challenging to design suitable tasks for dialogue summarization and find the best combination of tasks in different scenarios."
   - Sec. 6.3, Page 22 Line 7: "Additional data in previous work are summarized in Figure 7. These data are always used in the following ways: Pre-training with corresponding training objectives..... Mixing with dialogue summarization training data..... The advantages and disad-

vantages of using additional data are as follows: The language understanding ability among different corpora is the same intrinsically. As a result, additional data helps dialogue summarization, especially in low-resource settings, which further alleviates the burden of summary annotation by humans.....Training with more data is not always effective [126, 202], especially when the divergence between the additional corpus and original dialogue summarization corpus is huge. Elaborate data augmentation approaches avoid this problem when training data is not too scarce."

- Sec 8.1 Paper Analysis, Page 28: "The total number of papers on abstractive dialogue summarization investigated in this survey is 96. As shown in Figure 8, 33 of them propose new datasets and 67 make novel technical contributions. The other 9 papers are either a survey, a demo, or other strongly related papers.....There is no significant difference in the number of datasets between well-researched domains and the others. However, the release time and availability of different datasets vary.....It's a good sign that high-quality corpora, such as AMI and SAMSum, lead to a prosperous of techniques for dialogue summarization, but also raise a worry about the generalization ability of current techniques because of their over-reliance on specific datasets which may lead to over-fitting....."
- Sec 8.2 Future Directions, Page 30 Line 14: "We discuss some possible future directions and organize them into three dimensions: task scenarios, approaches and evaluations.....Multi-session dialogue summarization is required when conversations occur multiple times among the same group of speakers. The Information mentioned in previous sessions becomes their consensus and may not be explained again in the current session.....Generalizable and non-labored techniques have attracted increasing attention on other dialogue modeling tasks.....Compared with traditional dialogue summarization systems, LLM-based methods largely alleviate the tedious human labor and can be more generalizable due to the removal of unintended annotation artifacts. Nevertheless, approaches that are previously applied to small pre-trained language models in this survey may also provide inspirations and be adapted to augment LLMs for better dialogue summarization performance.....Factual errors caused by the mismatch between speakers and events are common as a result of complicated discourse relations among utterances in dialogues.....With the strong generation ability of current LLMs, there's also a doubt that whether the previous taxonomy of error types and evaluation metrics is still suitable. In a word, both meta-evaluation benchmarks and evaluation methods call for innovations."

2. Coverage of evaluation metrics. Compared with the introduction of methods and datasets, the organization and discussion on evaluation metrics is still relatively limited in this version. However, the evaluation metrics, especially automatic ones, are very important for the future development of dialogue summarization models. A well-expanded section 7.2 would make the survey more comprehensive. For example, adding summarization tables and in-depth discussions, like other sections, can also be helpful. Meantime, the ideal evaluation for ODS and TDS can be different considering their respective purposes. Here are some references in the aspect of evaluation (mostly on faithfulness):

   a) [NAACL 2021, for general abstractive summarization]: Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics.

   b) [NAACL 2022, for dialogue summarization, faithfulness]: CONFIT: Toward Faithful Dialogue Summarization with Linguistically-Informed Contrastive Fine-tuning

   c) [EMNLP 2022, for dialogue summarization faithfulness evaluation]: Analyzing and Evaluating Faithfulness in Dialogue Summarization.

d) [NAACL 2022, discussion on dialogue evaluation metrics]: DialSummEval: Revisiting Summarization Evaluation for Dialogues

AUTHORS' RESPONSE: We incorporated a table (Table 3) summarizing the evaluation metrics currently used for dialogue summarization, and explain the intuition behind different groups of metrics based on the task characteristics of ODS and TDS respectively. We also added some necessary citations in Sec. 7.2 "Evaluation Metrics" and Sec 8.2 "Future Directions" considering the limited research for dialogue summarization evaluation. Details are as follows:

- Table 3 for evaluation metrics, Page 27: Please refer to the manuscript.
- Sec. 7.2 intuition of ODS and TDS, Page 26 Line 32: "Instead of comparing only with the whole reference summary, most researches for TDS only consider key words/phrases while ignoring other common words for measuring the information coverage. In other words, evaluation for TDS emphasizes the coverage of key information which are generally domain-specific terms and can be easily recognized..... ODS pays less attention on information coverage due to the higher subjectivity on salient information selection. Instead, measuring the factual consistency of generations gains increasing attention.....Information correctness of the generated summary is also important for TDS. For instance,....."
- We cited b), c) and d) in Sec. 8.2.3:

  b), Page 18 Line 18: "Tang et al. [162] also designated summaries where negative summaries are constructed for different error types, such as swapping the nouns for wrong reference and object errors....."; "Tang et al. [162] also proposed circumstantial error, negation error, object error, tense error and modality error for more detailed scenarios. All of their error types can also be grouped into two classes, where the information missing and redundancy are for the coverage of key information, and the rest are for factual consistency....."; "Tang et al. [162] introduced a taxonomy of factual errors for abstractive summarization and did human evaluation based on this categorization without proposing new automatic metrics."

  b) and c), Page 33 Line 18: "Wang et al. [174] classified factual errors in a similar way to Tang et al. [162] and propose a model-level evaluation schema for discriminating better summarization models, which is different from the widely-accepted sample-level evaluation schema that scores generated summaries and can further scoring the model based on their corresponding outputs. They evaluated the model by calculating the generation probability of faithful and unfaithful summaries collected by rule-based transformations based on their taxonomy. The generalization ability for this work among different datasets and scenarios is doubtful, since a similar work for news summarization, FactCC [77], which is a metric trained based on rule-based synthetic datasets shows a poor generalization ability by Laban et al. [79]."

  d), Page 33 Line 1: "Recent work from Gao and Wan [54] re-evaluated 18 evaluation metrics and did a unified human evaluation with 14 dialogue summarization models on SAMSum dataset. Their results not only show the inconsistent performances of metrics between document summarization and dialogue summarization, and none of them excel in all dimensions for dialogue summarization, but also raise a warning on rethinking whether recently proposed complex models and fancy techniques truly improve the basic pre-trained language model....."
- We didn't cite the paper a), since this paper is a meta-evaluation benchmark with a taxonomy of factual error designed for document summarization, which is not related to dialogue summarization. Our survey cites other evaluation metrics for factual errors adopted in dialogue

summarization works:

Page 27 Line 4: "A QA-based model [173] is borrowed by Zhao et al. [204]. It follows the idea that factually consistent summaries and documents generate the same answers to a question. NLI-based methods [116] that require the content in the summary to be fully inferred from the dialogue were adopted by Liu et al. [106]. Liu and Chen [107] automatically evaluate inconsistency issues of person names by using noised reference summaries as negative samples and training a BERT-based binary classifier. Asi et al. [8] used the FactCC metric from Kryściński et al. [77] where the model was trained only with source documents with a series of rule-based transformations....."

3. Coverage of more traditional methods. Even though dialogue summarization methods are blooming only in recent years, there are some early works in this field. The current survey mainly focuses on the work after the proposal of neural encoder-decoder architecture. There is only a small paragraph in the introduction that explains some early efforts. To present a holistic review for readers with different backgrounds, it would be better to include the development roadmap from traditional to modern methods. Probably one subsection is enough.

   AUTHORS' RESPONSE: We added the development roadmap from traditional to modern methods at the beginning of Sec. 3 Overview of Approaches as follows:

   - Page 7 Line 2: "In abstractive text summarization, early researchers tried non-neural abstractive summarization methods [11], which used statistical models to recognize important words and sentences and then concatenate them into a final summary with or without pre-defined templates. The most direct way is to select a set of keywords from input [131], such as log-likelihood ratio test [93], which identified the set of words that appear in the input more often than in a background corpus. Another way is to assign weights to all words in the input. Most popular such work relied on TF-IDF weights [12]. Word weights have also been estimated by supervised approaches with typical features, including word probability and location of occurrence [154]. Some other traditional work directly focuses on predicting sentence importance, by either emphasizing select sentences that match the template of summaries or selecting the sentences in which keywords appeared near each other. Such sentences can better convey important information and be selected as a summary [18, 98]. Researchers also productively explored the relationship between word and sentence importance, and tried to estimate each in either supervised or unsupervised framework [101]. Since 2015, neural-based abstractive text summarization models [105, 129, 143, 146] began to be widely developed. These methods adapt recurrent neural network (RNN), convolutional neural network (CNN) and Transformer architecture for sentence representation. Benefiting from the semantic representation learned from neural networks and large training data, neural-based summarization methods outperform non-neural methods, especially in the aspect of fluency and semantic coherence."

4. Below are some recent work (may have come out after the authors submitted the manuscript) that the authors might want to include in their taxonomy and survey:

   - Picking the Underused Heads: A Network Pruning Perspective of Attention Head Selection for Fusing Dialogue Coreference Information, ICASSP 2023

   - Dynamic sliding window for meeting summarization, Interspeech 2022

   - Entity-based De-noising Modeling for Controllable Dialogue Summarization, SIGDIAL 2022

   AUTHORS' RESPONSE: We added these recent work in our survey at the following positions:

- Sec. 4.2.2, Page 15 Line 21: "The rest modify attention heads in the Transformer architecture with the constructed graphs from a model pruning perspective. Liu et al. [112] replace the attention heads that represent the most coreference information with their coreference graph, while Liu and Chen [110] replace the underused heads with similar coreference graphs."
- Sec. 4.2.1, Page 13 Line 16: ".....while Liu and Chen [109] adopted a dynamic stride size which predicts the stride size by generating the last covered utterance at the end of Y'."
- Sec. 5.2, Page 18 Line 29: "Word-level masks for ..... entities [70, 108],..... are considered in previous work, for a better understanding of the complicated speaker characteristics and capturing salient information. "

We appreciate the hard work by the editors and the reviewers, and hope that our revisions meet their requirements.

Once again, thank you very much for your comments and suggestions.

Sincerely,

Qi Jia, Yizhu Liu, Siyu Ren
and Kenny Q. Zhu