DialogSum Challenge: Summarizing Real-Life Scenario Dialogues

Yulong Chen[♠♡], Yang Liu[♣], Yue Zhang[♡◇]

[♠] Zhejiang University, China

[♡] School of Engineering, Westlake University, China

[♣] ILCC, School of Informatics, University of Edinburgh, UK

[◇] Institute of Advanced Technology, Westlake Institute for Advanced Study, China

yulongchen1010@gmail.com inf.yangl@outlook.com

yue.zhang@wias.org.cn

Abstract

We propose a shared task on summarizing real-life scenario dialogues, *DialogSum Challenge*, to encourage researchers to address challenges in dialogue summarization, which has been less studied by the summarization community. Real-life scenario dialogue summarization has a wide potential application prospect in chat-bot and personal assistant. It contains unique challenges such as special discourse structure, coreference, pragmatics and social common sense, which require specific representation learning technologies to deal with. We carefully annotate a large-scale dialogue summarization dataset based on multiple public dialogue corpus, opening the door to all kinds of summarization models.

Dialogue from DialogSum:
#Person_1#: Good morning. I wonder whether you have got an answer from your superior.
#Person_2#: Yes, we had a meting about it yesterday afternoon.
#Person_1#: What's the answer?
#Person_2#: We decided that we could agree to your price, but we are a bit worried about the slow delivery.
#Person_1#: Let me see. I quoted your delivery in three months, didn't I?
#Person_2#: Yes, but we hope that the wool could reach us as soon as possible.
#Person_1#: I thought you would. So I rang Auckland last night. As you are our biggest customer, they agreed to ship the order on the first vessel available that will leave Auckland next month.
#Person_2#: Good, if you agree we'll draft the agreement right away and sign it then.
#Person_1#: By all means.

Summary from DialogSum: #Person_1# and #Person_2# agree to sign an agreement ***since*** #Person_1# could speed up the delivery as #Person_2# hopes.

Figure 1: An example from DialogSum dataset.

1 Task Overview

The *DialogSum Challenge* asks a model to generate a salient, concise, fluent, and coherent summary, given a piece of multi-turn dialogue text. The dialogue summary is highly abstractive in nature, and is supposed to be objective compared with monologue summarization. We will conduct both automatic and manual blind evaluation on the submitted models. In particular, to address unique challenges in dialogue summarization, we will manually evaluate system-generated summaries from multiple aspects designed for dialogue summarization, including coreference information, discourse relation, intent identification and objective description. An example is shown in Figure 1, where the summary describes main events in a business conversation.

2 Motivation

Thanks to the advance in neural network models, and availability of large scale labeled datasets, recent research has achieved promising progress on summarizing monologic texts, such as news articles (Liu and Lapata, 2019; Gehrmann et al., 2018), patents (Pilault et al., 2020) and academic papers (Koncel-Kedziorski et al., 2019). However, dialogue, as an important channel for achieving communicative intents, differs from monologic texts in nature and has received significantly less attention from the summarization research community. A major reason is the paucity of suitable dialogue summarization datasets.

To cope with this problem, we build a large scale labeled summarization dataset for real-life scenario dialogues, DialogSum (Chen et al., 2021). An example from DialogSum is shown in Figure 1. Compared with existing dialogue summariztaion datasets (Carletta et al., 2005; Gliwa et al., 2019; Zhong et al., 2021; Zhu et al., 2021), DialogSum is useful for training neural models and is staying in the spoken domain as opposed to the written chat domain. Also, it contains diverse task-oriented dialogues that cover a wide range of daily-life topics. Summarizing those dialogues is useful for both business (e.g. help a business find common needs) and personal uses (e.g. track important events as

Proceedings of the 14th International Conference on Natural Language Generation (INLG), pages 308–313,
Aberdeen, Scotland, UK, 20-24 September 2021. ©2021 Association for Computational Linguistics

Dialogue from DIALOGSUM:
#Person_1#: Good morning. What can I do for you?
#Person_2#: I'm in Room 309. ***I'm checking out today.*** Can I have my bill now?
#Person_1#: Certainly. Please wait a moment. Here you are.
#Person_2#: Thanks. ***Wait…***What's this? The 30 dollar for?
#Person_1#: ***Excuse me…*** The charge for your laundry service on Nov. 20th.
#Person_2#: But I didn't take any laundry service during my stay here. ***I think you have added someone else's.***
#Person_1#: ***Ummm…*** Sorry, would you mind waiting a moment? We check it with the department concerned.
#Person_2#: No. As long as we get this straightened out.
#Person_1#: I'm very sorry. There has been a mistake. We'll corrected the bill. Please take a look.
#Person_2#: ***Okay, here you are.***
#Person_1#: Goodbye.

Summary from DIALOGSUM: #Person_2# is checking out and asks #Person1# for the bill. #Person1# gives #Person_2# a ***wrong*** bill at first then corrects it.

Figure 2: Selected case from DIALOGSUM dataset.

Dialogue from DIALOGSUM:
#Person_1#: Hey, don't I know you from somewhere?
#Person_2#: No, sorry. I don't think so.
#Person_1#: Didn't you use to work at Common Fitness Gym?
#Person_2#: No, I'm afraid I did not.
#Person_1#: Oh, but I know you from somewhere else. Did you use to work at the movie theater downtown? You did. Yes. It's you. I go there all the time and you always sell me popcorn and soda.
#Person_2#: No, that's not me either. Sorry, ma'am. Perhaps I look familiar to you, but ...
#Person_1#: No, I know you. I have met you before! Hold on. Let me think. This is driving me crazy. I know that we've talked before. Oh, I remember now. You work at the Whole Bean Cafe on the corner. It that right?
#Person_2#: No, wrong again. Sorry, ma'am, but I really have to get going.

Summary from DIALOGSUM: #Person_1# **thinks that #Person_1# knows #Person_2# somewhere**, but #Person_2# denies it.

Figure 3: Selected case from DIALOGSUM dataset.

personal assistants). Empirical study and analysis demonstrate challenges in real-life scenario dialogue summarization (Chen et al., 2021).

To highlight the challenges in dialogue summarization, we propose real-life scenario dialogue summarization challenge, *DialogSum Challenge*, to encourage researchers to investigate such problems. The evaluation for dialogue summarization contains both automatic evaluation, i.e. ROUGE score (Lin, 2004) and BERTScore (Zhang et al., 2019), and human evaluation from multiple aspects to address corresponding challenges (c.f. Section 2.1 and Section 3.3.2). For human evaluation, we anonymize the submitted models, and evaluate them on corresponding hidden sub-test sets to ensure the fairness.

2.1 Unique Challenges in DIALOGSUM

Although dialogue summarization is in line with the philosophy of monologue summarization, we find some unique challenges in dialogue summarization.

First, because of special linguistic phenomena, the dialogue on the source side differs from monologue. Dialogue information flow is intuitively reflected in the **dialogue discourse structures** (Wolf and Gibson, 2005), where two utterances can be closely related even there is a large distance between them. Such a phenomenon is common in procedures and negotiations. For example, in Figure 1, the penultimate utterance "... *we'll draft the agreement... and sign it...*" is actually replying to the third utterance "*What's the answer?*", between which the utterances can be viewed as negotiation process and conditions. Also, frequent **coreference** and **ellipsis** make dialogue difficult to understand (Grosz et al., 1995; Quan et al., 2019). For example, to generate "wrong" in the summary in Figure 2, the model needs to understand "*I think you have added someone else's (laundry service on my bill)*", where "*my bill*" refers to "*#Person_2#'s bill*". These linguistic phenomena make dialogues difficult to encode using ordinary representation learning techonologies (Chen et al., 2021).

Second, compared with monologic summarization, dialogues are summarized from an *observer's* perspective, which requires summary to be **objective**. For example, in Figure 3, #Person_1#'s statements are actually awaiting to be confirmed. Human annotators identified such situation and used objective language ("*#Person_1# thinks that #Person_1#...*") to describe those statements. Also, the process of *perspective shift* (from interlocutor to observer) intuitively leads to morphology and lexical changes (e.g. the expression of referents and third-person singular predicates) and syntax changes (e.g. using written languages to describe spoken dialogues).

Third, dialogue summarization goes beyond summarizing dialogue contents, but also dialogue actions at the **pragmatic** level. For example, in the summary in Figure 1, "*agree*" summarizes both actions of #Person_1# and #Person_2#; in the summary in Figure 2, "*gives*" summarizes a single dialogue action of #Person_1#; in the summary in Figure 3, "*thinks*" and "*denies*" summarize multiple dialogue actions of #Person_1# and #Person_2#, respectively. It requires models to not only summarize *what speakers are saying*, but also *what they are doing*.

3 Task Description

The task for participants is to provide a model that generates a summary given the input dialogue text. Both automatic and manual evaluation will be conducted to measure model performance.

3.1 Data

The participant of *DialogSum Challenge* can start immediately, as the DIALOGSUM dataset has been already public [1]. We collect 13,460 dialogue data for DIALOGSUM from three public dialogue corpora, namely Dailydialog (Li et al., 2017), DREAM (Sun et al., 2019) and MuTual (Cui et al., 2020), as well as an English speaking practice website. In term of size, DIALOGSUM is comparable with SAMSum while its average dialogue length is much longer than SAMSum, which comforts the purpose of summarization and is thus more challenging. The dialogue data cover a wide range of daily-life topics, including diverse task-oriented scenarios. We ask annotators to summarize the dialogue from an *observer's* perspective.

To ensure the annotation quality, each summary has been checked twice by different people, where the reward and punishment mechanism is included. We also sample and check the data after the second checking. When any error is found in the sampling checking, we ask annotators to repeat annotation and checking the annotation batch until no error can be found. To monitor the annotation and analyze inter-annotator agreement, we randomly select 500 dialogue, and ensure they are annotated and checked by different annotators. For each dialogue, we compare its summary and compute their pairwise ROUGE as shown in Table 1, which demonstrates our high annotation quality. Those 500 dialogues result in our test set. The public dataset consists of training (12,460), dev (500) and test (500) sets. For test set, we provide 3 references.

In addition to the public DIALOGSUM dataset, we build a ***hidden*** test set that consists of 100 dialogues and human annotated summaries. This ensures that participants will not be able to optimize their models against the hidden test set.

For the competition, participants can follow our data setting to train, develop and test their models on the public DIALOGSUM. Using external training data is allowed. For automatic evaluation, we will use both public and hidden test sets. For human evaluation, we will use the multiple subsets from Chen et al. (2021), which are collected from the test set, but not public.

Human Annotated Summary	R1	R2	RL
Summary1 to Summary2	52.90	26.01	50.42
Summary1 to Summary3	53.85	27.53	51.65
Summary2 to Summary3	53.30	26.61	50.44
Average	53.35	26.72	50.84

Table 1: ROUGE scores between three human annotated summaries in test set.

3.2 Protocol

Following previous work (Syed et al., 2018, 2020), we divide the competition into three phrases: (1) participants will train proposed summarization models using the provided training data on their hardware; (2) after submission system opens, participants will make their trained model submission to the TIRA. When the test data is available on the system, it will automatically make blind evaluation on the submitted model; (3) after the submission deadline is due, we will start to evaluate summaries generated by the final submitted models via crowdsourcing workers from multiple aspects.

We plan the following schedule for *DialogSum Challenge*. Please note that dates may be modified when we know the detailed schedule of INLG 2022.

- **20th September, 2021:** The shared task announced along with data available; call for participants.

- **20th Dec, 2021:** The submission system and public leaderboard open; participants can submit trained models to the TIRA infrastructure; the TIRA infrastructure will automatically evaluate submitted models with automatic metrics; the online leaderboard will keep updating the best performance on both public test set and hidden test set.

- **20th Feb, 2022:** The deadline for final model submissions; manual evaluation via crowdsourcing begins.

3.3 Evaluation

Our evaluation contains both automatic evaluation metric and human evaluation.

[1] https://github.com/cylnlp/DialogSum

3.3.1 Automatic Evaluation

We will report ROUGE and BERTScore (Zhang et al., 2019). ROUGE measures the overlap of n-grams in the generated summary against the reference summary, intuitively reflecting model's capturing ability of salient information. We will use ROUGE as the main automatic evaluation metric. BERTScore computes a similarity score between the generated summary and reference summary on token level using contextual embeddings, which provides a more robust evaluation method for generation tasks. We will use BERTScore as a supplementary metric. We will report the lowest, highest and averaged scores on our multi-reference test set, to better evaluate model performance, including their variance.

3.3.2 Human Evaluation

We previsouly show that, although models can achieve high ROUGE scores, their generated summaries can contain many errors regarding dialogue understanding. Thus, we design human evaluation from multiple aspect based on Chen et al. (2021). To ensure the fairness, we will conduct human evaluation via Amazon Mechanical Turk, and each generated summary will be judged by three annotators to ensure the accuracy. All human annatators will read system-generated summaries and rate them based on following criteria.

Standard Summarization Metrics: Fluency, Consistency, Relevance and Coherence Following Kryscinski et al. (2019, 2020), we evaluate system-generated summaries from four dimensions, which have been widely used as standard summary evaluation criteria in human evaluation for monologue text. Human annotators will follow Kryscinski et al. (2019)'s criteria, and evaluate on a 50 randomly selected sub-testset.

Coreference Information Chen et al. (2021) find that a big challenge in dialogue summarization is that, because of interactive information flow, models show poor performance on correctly aligning interlocutors and their conversation actions/contents. Thus, we will ask human annotators to follow Chen et al. (2021)'s criteria and rate the summary on a 50 randomly selected sub-testset.

Intent Identification A comprehensive dialogue summary expresses interlocutors' intents (i.e. the function of their utterances), which is frequent in dialogues and essential to understanding dialogues. However, system-generated summaries usually focus on the consequence of a dialogue, and fail to correctly identify interlocutors' intents. Therefore, we will conduct human evaluation on intent identification on the 50 randomly selected sub-testset following Chen et al. (2021).

Discourse Relation Coherent summaries convey important relations between main events, and identifying discourse relations and using proper phrases to express them can be challenging for summarization systems (Xu et al., 2020). However, causally related events are usually not explicitly expressed, and the distance between them is long due to the unique dialogue discourse structure (Grosz et al., 1995). To quantify such challenge, we will conduct human evaluation on discourse relation following (Chen et al., 2021) on the discourse sub-testset.

Objective Description In addition to the above evaluation aspects, we also find that models tend to take all interlocutors' contents as ground truth while failing to reason whether their statements are just subjective assumptions or even defended to be fake. Therefore, we will evaluate whether system-generated summaries use objective language to describe dialogues, and give scores from -1, 0, 1, where 1 means all correct, 0 means partially correct and -1 means all incorrect.

Overview Score To give an overview score for each model, we will ask annotators to evaluate each summary along with the above multi-aspect evaluation scores and give a score from 1 to 5. The higher, the better.

3.3.3 Overview Ranking

As mentioned, models that achieve high performance regarding automatic evaluation still contain many errors. Thus, the final ranking will be determined by human annotators' judgements. However, as our human evaluation contains multiple aspects and the cost can be high, we will only conduct human evaluation on a limited number of candidate models, which show leading performance on automatic evaluation metrics against the hidden test set. Up to the top twenty submission will be considered as candidate model for human evaluations.

4 Conclusion

Different from existing summarization datasets, the DIALOGSUM poses unique challenges in dialogue summarization. And we believe that *DialogSum*

Challenge will open new avenues for researchers to investigate solutions and study the linguistic phenomena in dialogue summarization.

5 Ethics Consideration

Dialogue data of *DialogSum Challenge* are collected from DailyDialog, DREAM, MuTual and an English practicing website that all are public to academic use and do not contain any personal sensitive information.

The construction of additional *DialogSum Challenge* hidden test set involves manual annotation. We ask annotators to write summarize limited to given dialogues, thus no personal or sensitive information is introduced.

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.

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

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. MuTual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1406–1416, Online. Association for Computational Linguistics.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 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, Hong Kong, China. Association for Computational Linguistics.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. 2019. Text Generation from Knowledge Graphs with Graph Transformers. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2284–2293, Minneapolis, Minnesota. 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 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 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 the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081, Florence, Italy. Association for Computational Linguistics.

Jonathan Pilault, Raymond Li, Sandeep Subramanian, and Chris Pal. 2020. On extractive and abstractive neural document summarization with transformer language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9308–9319, Online. Association for Computational Linguistics.

Jun Quan, Deyi Xiong, Bonnie Webber, and Changjian Hu. 2019. GECOR: An end-to-end generative ellipsis and co-reference resolution model for task-oriented dialogue. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4547–4557, Hong Kong, China. Association for Computational Linguistics.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Shahbaz Syed, Wei-Fan Chen, Matthias Hagen, Benno Stein, Henning Wachsmuth, and Martin Potthast. 2020. Task proposal: Abstractive snippet generation for web pages. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 237–241.

Shahbaz Syed, Michael Völske, Martin Potthast, Nedim Lipka, Benno Stein, and Hinrich Schütze. 2018. Task proposal: The tl; dr challenge. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 318–321.

Florian Wolf and Edward Gibson. 2005. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–287.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, et al. 2021. Qmsum: A new benchmark for query-based multi-domain meeting summarization. *arXiv preprint arXiv:2104.05938*.

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