DIALOGSUM: A Real-Life Scenario Dialogue Summarization Dataset

Yulong Chen[♠♡], **Yang Liu**[♣], **Liang Chen**[♦], **Yue Zhang**[♡◇]
[♠] Zhejiang University
[♡] School of Engineering, Westlake University
[♣] Microsoft Cognitive Services Research
[♦] College of Software, Jilin University
[◇] Institute of Advanced Technology, Westlake Institute for Advanced Study
yulongchen1010@gmail.com *yaliu10@microsoft.com*
chenliang5518@mails.jlu.edu.cn *yue.zhang@wias.org.cn*

Abstract

Proposal of large-scale datasets has facilitated research on deep neural models for news summarization. Deep learning can also be potentially useful for spoken dialogue summarization, which can benefit a range of real-life scenarios including customer service management and medication tracking. To this end, we propose DIALOGSUM, a large-scale labeled dialogue summarization dataset. We conduct empirical analysis on DIALOGSUM using state-of-the-art neural summarizers. Experimental results show unique challenges in dialogue summarization, such as spoken terms, special discourse structures, coreferences and ellipsis, pragmatics and social common sense, which require specific representation learning technologies to better deal with.

(a) 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.

(b) Dialogue from SAMSum:
…
Leo: BTW what are those pics?
Ryan: Pics from Italy!!! :):):):)))))))))
Leo: Yeah. They seem nice. (‘A`)
Ryan: That's all???? I need more reactions!!!!!!!!!!
Leo: I'm tied to this office and working like a slave. AM I SUPPOSED TO SAY \"I AM SO JEALOUS!!!!!!!!\"?😬😬😬
…

Summary from SAMSum: Ryan is in Italy while Leo is working hard and wishing he could win the lottery.

Figure 1: An example from DIALOGSUM dataset compared with an example from SAMSum dataset.

1 Introduction

Text summarization is the task of automatically generating a concise, salient, coherent and fluent summary of a given set of documents (Radev et al., 2002). Thanks to the advance in neural network models and the availability of large-scale labeled datasets, recent research has achieved promising progress on summarizing monologic texts such as news articles (Paulus et al., 2018; Gehrmann et al., 2018; Liu and Lapata, 2019; Liu et al., 2020), patents (Pilault et al., 2020) and academic papers (Koncel-Kedziorski et al., 2019).

However, dialogue, as an important channel for achieving communicative intents (Bender and Koller, 2020), has received significantly less attention from the summarization research community. One main reason is the paucity of a suitable summarization dataset built on dialogue texts. Most existing research uses the AMI meeting corpus (Carletta et al., 2005), which consists of 137 dialogues obtained from virtual multi-party meeting recordings. However, research on the corpus is limited to its small scale. SAMSum (Gliwa et al., 2019) is a recently released *written* online dialogue summarization dataset, which contains 16k online chats with corresponding summaries. However, it focuses on conversations via messenger apps, which are rather short (around 94 tokens per conversation) and their language style and topics also differ from *spoken daily dialogues*.

A comparison between the real-life scenario dialogue and online chat is shown in Figure 1. Online-chat messages contain unique tokens (e.g., “BTW”), emoticons (e.g., “:)”) and emojis (e.g., “😬”). In contrast, daily conversations have a different

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 5062–5074
August 1–6, 2021. ©2021 Association for Computational Linguistics

Datasets	Lan. style	Domain	Scenario	Dialogs	Data size	#Tokens/dial.	#Tokens/turn	#Comp. rate
AMI	spoken	single	meeting	137	100hrs (video)	4,757	16.5	0.07
SAMSum	written	multiple	online	16,369	1.5M (token)	94	8.4	0.30
DIALOGSUM	spoken	multiple	daily life	13,460	1.8M (token)	131	13.8	0.18

Table 1: Comparison between DIALOGSUM and other public dialogue summarization datasets. Lan. stands for language. Dial. stands for dialogue. # stands for the average result. Comp. stands for compression. The compression rate is the ratio of the length of the summary divided by the length of the original text.

and more formal style. In addition, real-life dialogues have more diverse task-oriented scenarios and topics compared to online chit-chats. For example, online-chat messages in SAMSum are about leisure and social chats, but real-life dialogues contain business negotiation (Figure 1(a)). Intuitively, automatically summarizing such dialogues can help a business find common needs or complaints from customers. With the rise of personal assisting chatbots, summarizing dialogues from different aspects of daily life can also be useful for personal record management and other applications.

We introduce Real-Life Scenario Dialogue Summarization (DIALOGSUM), a large-scale summarization dataset for dialogues. Dialogue data for DIALOGSUM are collected 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. These datasets contain face-to-face spoken dialogues that cover a wide range of daily-life topics, including schooling, work, medication, shopping, leisure, travel. Most conversations take place between friends, colleagues, and between service providers and customers. We clean and preprocess the dialogue data into a unified format, and ask annotators to summarize them from an observer perspective. Topics are also manually labeled for each dialogue. An example of DIALOGSUM is shown in Figure 1(a), where the summary expresses the main content in a business conversation.

The contribution of DIALOGSUM can be stated from two perspectives. First, from the perspective of downstream applications, summarizing daily spoken dialogues can be useful for both business and personal uses. Dialogue summaries can also be useful for personal assistants to keep track of important events as such business negotiation. Second, from the method perspective, DIALOGSUM has a larger scale of long dialogue data, which can facilitate the study of dialogue summarization using neural network models. The number of dialogues in DIALOGSUM is orders of magnitude larger than in AMI, which can be useful for training large neural network models for dialogue summarization. The average length of dialogues in DIALOGSUM is 39.8% longer than in SAMSum. To our knowledge, we are the first to release a large-scale real-life scenario dialogue summarization dataset.

We empirically investigate the performance of state-of-the-art neural summarization models on DIALOGSUM, comparing the characteristics of the spoken daily dialogue summarization dataset with standard news summarization benchmarks and the online chat summarization benchmark SAMSum. Experimental results show that DIALOGSUM is more amenable to abstractive summarizers, while being relatively more challenging compared to the existing summarization datasets. We find that main difficulties arise from discourse structures in multi-turn dialogues, as well as the need for book-keeping both entities and events mentioned in turns of utterances. We release our dataset at https://github.com/cylnlp/DialogSum.

2 The DIALOGSUM Dataset

2.1 Dialogue Data Preparation

Data Collection DailyDialog is a dataset consisting of 13k multi-turn dialogues, obtained from websites that aim to help English learners to practice English speaking. DREAM and MuTual are dialogue understanding datasets, consisting of 6k and 9k speech transcripts, respectively, both collected from online English listening exam materials. In order to further increase the diversity of data, we crawl additional dialogues from another English speaking practice website[1] which aims to provide English learners with conversation examples in real life practical circumstances, such as business negotiation and banking services.

Although dialogues of DIALOGSUM are from different sources, they all share important characteristics that are in line with what we expect. First, as mentioned earlier, these dialogues are under rich

[1] http://www.tingroom.com

Source	Num. dial.	Tokens	% in DIALOGSUM
DailyDialog	7837	980,398	58.22
DREAM	2028	301,098	16.94
MuTual	1870	260,139	13.89
Crawled	1725	238,902	12.82

Table 2: Proportions of dialogue sources in DIALOGSUM.

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

real-life scenarios. Unlike chitchats, these conversations have clear communication patterns and intents, making them more suitable and valuable to serve as summarization sources (Carletta et al., 2005). Moreover, their multi-turn dialogue lengths are within a reasonable scale and are longer than chitchats[2], which comforts the purpose of automatic summarization. Greater lengths also indicate these dialogues contain more events and discourse relations between them. Properly selecting vital events and identifying their relations make summarizing these dialogues more challenging.

Data Cleaning and Pre-Processing We delete non-English characters, correct typos and grammatical errors, and further filter out duplicated data based on text similarity. After deduplicating, proportions of the data sources are summarized in Table 2. Because of different data processing methods and annotation procedures, original dialogues in DailyDialog, DREAM and MuTual are in different formats. We follow previous work (Li et al., 2017; Zhang et al., 2018; Budzianowski et al., 2018; Dinan et al., 2019) and preprocess them into a bi-turn dialogue flow, merging continuous turns of the same speaker into one utterance. Also, we add tags (e.g. `#Person_1#` and `#Person_2#` in Figure 1(a)) before each dialogue turn, to distinguish speakers. The final DIALOGSUM dataset contains 13,460 dialogues, which are divided into training (12,460), validation (500) and test (500) sets.

2.2 Annotation

We ask annotators to write dialogue summaries based on following criteria: the summary should (1) convey the most salient information of the dialogue and; (2) be brief (no longer than 20% of the conversation length) and; (3) preserve important named entities within the conversation and; (4) be written from an observer perspective and; (5) be written in formal language.

We require our annotators to pay extra attention to the following aspects.

Tense Consistency: Annotators should take the moment that the conversation occurs as the *present* time, and choose a proper tense to describe events *before* and *after* the ongoing conversation.

Discourse Relation: If summarized events hold important discourse relations, particularly causal relation, annotators should preserve the relations if they are also in the summary.

Emotion: Different from newspaper and academic articles, social conversations in DIALOGSUM are often implied with emotions. Therefore, we ask annotators to explicitly describe important emotions related to events in the summary.

Intent Identification: Rather than merely summarizing the consequences of dialogues, annotators should also describe speakers' intents in summaries, if they can be clearly identified.

In addition to the above, annotators should use person tags to refer to different speakers if real names cannot be detected from the conversation. Annotators are also asked to write a short (around 3 tokens) topic for each dialogue. Appendix A shows the list of topics.

2.3 Quality Control

To ensure quality, before formal annotation, we ask annotators to annotate training samples until they pass our examination and meet our requirements. After annotation, we check summaries by cross-validation between different annotators twice. During the checking process, bonus is paid to checkers who find unqualified summaries, and penalty is given to annotators whose annotation is found with mistakes. In case of appeal, we make the final decision. After the second checking, we sample 10% summaries and manually check the samples ourselves. If errors are found in an annotation batch, we ask corresponding annotators to self-check and re-annotate the whole batch and repeat this checking and sampling processes.

To further control the quality, and to analyze inter-annotator agreement, for each dialogue in the

[2] The average numbers of tokens for multi-turn dialogues are: Dailydialog: 118.8, DREAM: 124.6, MuTual: 136.1.

Dataset	% of novel n-grams				Lead			Longest			Ext-Oracle		
	unigram	bigram	trigram	4-gram	R1	R2	RL	R1	R2	RL	R1	R2	RL
CNN	16.75	54.33	72.42	80.37	29.15	11.13	25.95	-	-	-	50.38	28.55	46.58
DailyMail	17.03	53.78	72.14	80.28	40.68	18.36	37.25	-	-	-	55.12	30.55	51.24
NY Times	22.64	55.59	71.93	80.16	31.85	15.86	23.75	-	-	-	52.08	31.5	46.72
XSum	**35.76**	**83.45**	**95.50**	**98.49**	**16.30**	**1.61**	**11.95**	-	-	-	**29.79**	**8.81**	**22.65**
SAMSum	32.63	77.22	89.27	94.83	31.41	8.70	30.41	32.13	10.13	29.11	44.60	17.37	39.38
DialogSum	26.28	76.94	89.16	94.53	27.52	6.78	27.31	24.15	6.25	22.73	37.90	13.88	34.04

Table 4: Corpora statistics and extractive methods on CNN/DailyMail, NY Times, XSum, SAMSum and DialogSum. Part of results is from Narayan et al. (2018). All results are computed on test sets. For DialogSum, the results are the average of multi-reference results.

test set, we provide three summaries written and checked by different annotators. For each test dialogue, we compare its and compute their pair-wise ROUGE (Lin, 2004) scores. Table 3 reports their averaged F_1 scores of ROUGE-1 (R1), ROUGE-2 (R2) and ROUGE-L (RL). We see R2 is relatively low while RL is high, which suggests that annotators' usage of language is variable, but the main content and logical order are mostly the same.

2.4 Characteristics of DialogSum

We empirically compare DialogSum with existing news summarization datasets and SAMSum. CNN/DailyMail (Hermann et al., 2015), NY Times (Sandhaus, 2008) and XSum (Narayan et al., 2018) are large-scale summarization datasets from the news domain, written in a monologic structure. XSum is a dataset designed specifically for abstractive summarization.

First, we compare the percentages of novel n-grams in the reference summary against the source document/dialogue. This intuitively reflects the level of abstraction of annotated summaries. As shown in Table 4, except for XSum, which is designed to be highly abstractive, dialogue-based summarization datasets contain more novel n-grams in the summaries. We also find that the percentage of novel unigrams in DialogSum is 26%, 6% lower than in SAMSum, but novel bigrams, trigrams and 4-grams are about the same as SAMSum. We believe that the relatively lower novel unigram proportion in DialogSum compared to SAMSum is because of our pre-processing and annotation criteria. SAMSum's summaries include real names, third-person singular pronouns, which can be diverse across the dialogues. In contrast, DialogSum uses tags such as `#Person_1#` to refer to persons whatever they are subjective, objective, or possessive. This constrains the proportion of novel unigrams to be lower.

Second, we compare the datasets using several extractive summarization methods. Following previous summarization work (Liu, 2019; Pilault et al., 2020), we report R1, R2 and RL here. Lead creates summaries by selecting the first n sentences from source texts. Longest is designed for dialogue summarization (Gliwa et al., 2019). It selects the n longest utterances as a summary, which gives better ROUGE scores than Lead on SAMSum. Ext-Oracle creates summaries by choosing n sentences that have the highest ROUGE against reference summaries. It can be viewed as an upper bound for extractive summarization. We report results of Lead-3, Longest-3 and Ext-Oracle-2 on SAMSum, and Lead-2, Longest-2 and Ext-Oracle-2 on DialogSum, where n is searched for each dataset in range of 1 to 6.

The results are shown in Table 4. In terms of Lead, DialogSum sees the lowest R1 and R2 except for XSum, showing that it is in nature a highly abstractive summarization dataset. SAMSum is less abstractive than DialogSum by all ROUGE scores, which is likely because the compression rate of SAMSum (0.30) is higher than DialogSum (0.18) (Table 1). The higher compression ratio suggests the summary contains denser information in the original text. The same conclusion can be found by using the Longest method. By using the Ext-Oracle method, we find that DialogSum is the most challenging dataset for extractive summarizers except for XSum, which is carefully designed for evaluating abstractive summarizers.

3 Experiments

We experiment with several abstractive summarization baselines to further understand the characteristics and challenges of DialogSum. Following Gliwa et al. (2019), we concatenate utterances of a dialogue as the input. For pretrained models, we only finetune them on corresponding datasets.

Model	CNNDM			XSum			SAMSum			DialogSum		
	R1	R2	RL	R1	R2	RL	R1	R2	RL	R1	R2	RL
Transformer	40.21	17.76	37.09	29.41	9.77	23.01	37.20	10.86	34.69	35.91	8.74	33.50
UniLMv2$_{BASE}$	43.16	20.42	40.14	44.00	21.11	36.08	50.53	26.62	48.81	47.04	21.13	45.04
BART$_{LARGE}$	**44.16**	**21.28**	**40.90**	**45.14**	**22.27**	**37.25**	**53.12**	**27.95**	**49.15**	47.28	21.18	44.83

Table 5: Results of abstractive models on CNNDM, XSum, SAMSum and DialogSum. For DialogSum, we give the average of multi-reference results.

3.1 Models

Transformer We take Transformer (Vaswani et al., 2017) as a non-pretrained abstractive baseline. For dialogue summarization, we follow Gliwa et al. (2019), using the same hyper-parameters for news summarization[3], but changing the minimum length to 15. We train the 6-layer Transformer model with Adam (Kingma and Ba, 2014) for 100,000 steps. Copy attention mechanism is applied and the dropout rate is set to 0.1.

UniLMv2 UniLMv2 (Bao et al., 2020) is a recently released pretrained language model for autoencoding and partially autoregressive language modeling. Here we use UniLMv2$_{BASE}$ as a strong abstractive model. For dialogue summarization, we train the model with Adam for 100,000 steps with 2,000 warmup steps and learning rate is set to $1.5e^{-5}$.

BART BART (Lewis et al., 2020) is an encoder-decoder Transformer model pretrained on a large corpus using a denoising autoencoder task. We use the large version of BART and finetune it with 5,000 training steps/200 warmup steps for dialogue summarization. Learning rate is set to $3e^{-5}$.

3.2 Results

Table 5 presents the experimental results. In general, we find that non-pretrained abstractive models outperform LEAD (Table 4), and the best results are achieved by pretrained models, despite the fact that BART$_{LARGE}$ and UniLMv2$_{BASE}$ are pretrained on monologic texts.

Extractive Summary vs Abstractive Summary Transformer gives similar results on CNNDM and better results on XSum, SAMSum and DialogSum compared to LEAD, and pretrained models show better performance than EXT-ORACLE on all datasets except for CNNDM. In particular, pretrained models outperform Transformer by $13.07 \sim 14.24\%$ RL on XSum, $14.12 \sim 14.46\%$ RL on SAMSum, and $11.33 \sim 11.54\%$ on DialogSum, while only $3.05 \sim 3.81\%$ on CNNDM. We believe that it is because CNNDM is a highly extractive dataset (Section 2.4). The key to summarizing CNNDM is to correctly understand intersentence relations within long documents, and extract important sentences. In contrast, XSum, SAMSum and DialogSum are more abstractive, which require a model to paraphrase. And the strong generation capability of pretrained models can bring great improvements on them. We also see that, for abstractive datasets, model performance decreases as document length grows (Avg. length: SAMSum - 93.8, DialogSum - 131.1, Xsum - 431.1) and compression rate decreases (Comp. rate: SAMSum - 0.30, DialogSum - 0.18, XSum - 0.05). This explains why SAMSum is the easiest dataset.

Spoken vs Written All three models perform better on dialogue summarization datasets, compared with XSum. This can be potentially because XSum is naturally highly abstractive, and thus more challenging. We also compare improvement brought by pretrained models that are trained on large written texts.

Still in Table 5, the improvement on DialogSum is the least. BART$_{LARGE}$ outperform Transformer by 15.73% R1 on XSum, 15.92% R1 on SAMSum, but 11.37% R1 on DialogSum. It demonstrates that SAMSum has overall more written style than DialogSum, and also suggests that dialogue and monologue are different. This can be explained by the design of written app-chat annotation in SAMSum (Gliwa et al., 2019).

DialogSum vs SAMSum As shown in Table 5, model performance is steadily lower on DialogSum than SAMSum. As stated, DialogSum is more abstractive, open-domain, and spoken analogous. One more possible reason for the lower performance on DialogSum is the longer input size. To better quantify the difference between these two dialogue summarization datasets, we further evaluate Transformer trained on DialogSum when tested on the SAMSum, and vice versa. As

[3] https://opennmt.net/OpenNMT-py/examples/Summarization.html

Trans. Test	R1	R2	RL
S2D	31.72(-5.48)	6.25(-4.61)	29.72(-4.97)
D2S	31.74(-4.17)	5.93(-2.81)	29.79(-3.71)

Table 6: Difference between DIALOGSUM (D) and SAMSum (S). Trans. stands for Transferred.

Summary	Fluency	Cons.	Relevance	Coherence
Summary 1	5	5	4.96	5
Summary 2	5	5	4.98	5
Summary 3	5	5	5	5
Avg.	5	5	4.98	5
Transformer	4	2.08	2.3	3.84
UNILMV2$_{\text{BASE}}$	4.8	3.84	4.06	4.34

Table 7: Human evaluation on human annotated summaries and model generated Summaries. Cons. stands for Consistency. Summary 1 - Summary 3 correspond to three summaries of a dialogue.

shown in Table 6, the performance of Transformer drops greatly when traiend on DIALOGSUM and tested on SAMSum, and vice versa. This shows that the two datasets have substantial differences. In addition, Transformer trained on DIALOGSUM performs better than on SAMSum, and shows lower performance drop, suggesting that DIALOGSUM can provide more generalization ability for training dialogue summarization models.

4 Human Evaluation

To better understand DIALOGSUM, we take a deeper investigation into the outputs of Transformer and UNILMV2 on DIALOGSUM by conducting human evaluation from multiple aspects.

Fluency, Consistency, Relevance and Coherence First, following Kryscinski et al. (2019, 2020), we implement human evaluation from four dimensions. *Fluency* evaluates the quality of individual generated sentences, *Consistency* evaluates the factual alignment between the source text and generated summary, *Relevance* evaluates the importance of summary content, and *Coherence* evaluates the collective quality of all sentences.

We randomly select 50 dialogues and their summaries from DIALOGSUM test, and ask a judge to give scores in scale from 1 to 5 along the four mentioned dimensions. The higher, the better. The judge also gives scores to human-annotated summaries to evaluate their quality. As shown in Table 7, human annotated summaries receive the best scores from all dimensions. UNILMV2$_{\text{BASE}}$ has steadily better scores than Transformer, but lower than human. Model-generated summaries have the highest scores on Fluency, while lowest on Consistency. It suggests that although model-generated summaries are grammatical and fluent, they still contain factual errors.

Model	Human Scores				ROUGE Scores		
	-1	0	1	Avg.	R1	R2	RL
Transformer	80%	17%	3%	-0.77	34.35	7.01	31.13
UNILMV2	43%	37%	20%	-0.23	43.78	17.91	40.97

Table 8: Human evaluation on discourse relations, with corresponding ROUGE scores on the sub-test set. Avg. stands for the averaged score here.

Discourse Relation Reasonable summaries should convey important relations between main events, and identifying discourse relations and using proper phrases to express them in summaries can be challenging for summarization systems (Xu et al., 2020). Take Figure 1 (a) for example, the human annotated summary connects two main events (underlined) using "*since*" to express their causal relation explicitly. However, the causal relation between those two events are not explicitly expressed in the dialogue, and the distance between them is long. Multiple turns usually correspond to more complicated discourse structure and relation. Also, similar with Chen and Yang (2020), we find that model performance decreases when the number of dialogue turns grows (See Appendix B).

To better evaluate model ability to disambiguate discourse relations in DIALOGSUM, we first collect discourse connectives from Penn Discourse Treebank (Miltsakaki et al., 2004), and check whether these connectives are included in summaries in the testset. If the three reference summaries of a dialogue all contain connectives, we assume that the dialogues have strong discourse signals. We choose 70 dialogues from DIALOGSUM in this way.

We then ask linguists who specialize in discourse to evaluate model outputs and give scores from $\{-1, 0, 1\}$, where 1 means that the generated descriptions of main events are reasonable and contain correct discourse connectives, 0 means that the descriptions are good but contain no discourse connectives and -1 means that the description is either incorrect or contains incorrect connectives. We ask the linguists to focus only on clauses or phrases that are essential to discourse relations, and ignore syntactic errors. We report the distribution of annotated scores in Table 8.

We can see that the most summaries generated by Transformer are scored as -1, and their average score is -0.77, close to -1. This means that Transformer is not only incapable of identifying discourse relations but also incapable of generating the main events correctly. UniLMv2 has a relatively smooth distribution over three categories and a better average score of -0.23, which is closer to 0, suggesting that UniLMv2 can mostly choose important events amongst the conversation. But the -1 still holds most proportion and its average result is still far from 1, indicating its incapability of understanding relations between events.

Compared to the full test set, the model performance on this sub-set generally decreases ($1.56 \sim 3.26\%$ lower of R1, $1.73 \sim 3.22\%$ of R2, $2.37 \sim 4.07\%$ of RL), which also suggests complicated discourse relations between events make summarization more difficult. The results indicate that further research is necessary for better representing dialogue discourse structures in order to obtain more reliable summarization systems.

Coreference Information To evaluate model's ability to distinguish different interlocutors, we ask a judge to evaluate whether interlocutors' names and their conversation actions/contents are correctly associated in the 50 randomly selected data, and give scores from $\{-1, 0, 1\}$, where 1 means that all names and actions/content in the summary are associated correctly, 0 means partial incorrectly, and -1 means all incorrectly. Here, we only focus on coreference information in generated summaries, and ignore other errors, such as incorrect syntax or failing to summarize salient information.

We report the distribution of annotated scores in Table 9. Most Transformer generated summaries are annotated as -1 and the average result is close to -1, suggesting that Transformer cannot generate clauses that express the same relation between arguments and predicates in original dialogues. The UniLMv2$_{\text{BASE}}$ has more 0-scored summaries, and the result is much higher, yet closer to 0, which indicates that although UniLMv2$_{\text{BASE}}$ can generate summaries containing correct clauses, but still have much inconsistency. The performance of both models indicates that Transformer is only capable of extracting important word-level information from dialogues in DialogSum, while UniLMv2$_{\text{BASE}}$ shows better performance on clause-level — it can identify the speakers and partially preserve coreference information, consistent with findings of Levesque et al. (2012) that pretraining is useful for coreference resolution. However, it is far from human annotations.

Model	**Human Scores**				**ROUGE Scores**		
	-1	0	1	Avg.	R1	R2	RL
Transformer	66%	28%	6%	-0.6	35.68	8.49	32.77
UniLMv2	4%	56%	40%	0.36	47.46	21.33	44.93

Table 9: Human evaluation on models' ability of preserving coreference information on DialogSum, with corresponding ROUGE scores. Avg. stands for the averaged score here.

Summary	**Human Scores**	
	-1	1
Summary 1	7.7%	92.3%
Summary 2	20.5%	79.5%
Summary 3	17.9%	82.1%
Avg.	15.4%	84.6%
Transformer	84.6%	15.4%
UniLMv2	30.8%	69.2%

Table 10: Human evaluation on models' ability of identifying interlocutors' intents.

Intent Identification As stated in Section 2.2, we ask annotators to include important intents of interlocutors in their summaries, addition to the consequences of dialogues. The intent here refers to the motivation of a speaker to initiate a conversation, e.g. "*want to do an annual physical*" (c.f. Figure 2, Dialogue-A). This can make summaries more comprehensive and readable. Therefore, we conduct corresponding human evaluation on whether interlocutors' intents are described in summaries in the 50 randomly selected data.

We first ask a judge to evaluate whether the intent is important to a dialogue, and we select 39 dialogues that contain important intents. Then, we ask the judge to give scores from $\{-1, 1\}$, where 1 means that intents are identified correctly, -1 means incorrectly. Note that we only focus on intent identification in the summary, and other errors should be ignored. We also ask the judge to evaluate human annotated summaries.

The distribution of annotated scores is shown in Table 10. We see that most summaries generated by Transformer are scored as -1, which means that Transformer is incapable of generating summaries that correctly convey speakers' intents. UniLMv2$_{\text{BASE}}$ shows much better performance, however, it is still below human performance.

5 Challenges in DialogSum

Compared to written texts, spoken dialogues can be more difficult for models to understand, and

DIALOGUE - A:	**DIALOGUE - B:**
#Person_1#: Hello, so how are we feeling today? **#Person_2#:** Things are going well for me, doctor. **#Person_1#:** Am I correct in thinking that you are here for your annual physical? **#Person_2#:** Yes, I am applying for new health insurance, and I need a physical examination to qualify. **#Person_1#:** Your basic physical exam will include lungs, heart, blood levels, and eyes, ears, and nose. **#Person_2#:** I've been having a little trouble breathing. Would you look into that, please? **#Person_1#:** We can do an allergy test, and later I can send you for an asthma test. **#Person_2#:** I would appreciate it. When you give me a blood test, what are you looking for? **#Person_1#:** I am going to check your cholesterol, blood sugar, and white blood cell count. **#Person_2#:** I am expecting the tests to go well. I have been taking good care of myself.	**#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 – A1: #Person_2# wants to do an annual physical examination to apply for new health insurance and says #Person_2#'s breathing is not good. #Person_1# explains the items and will do tests on #Person_2#'s breathing. **SUMMARY – A2:** #Person_1# explains the checking items in #Person_2#'s annual physical examination and will do test to look into #Person_2's breathing. **SUMMARY – A3:** #Person_2# is going through an annual physical examination to apply for new health insurance, and #Person_2# asks #Person_1# to look into the breathing.	**SUMMARY – B1:** #Person_2# is checking out and asks #Person1# for the bill. #Person1# gives #Person_2# a wrong bill at first then corrects it. **SUMMARY – B2:** #Person_1# helps #Person_2# correct a mischarged bill on laundry service and helps #Person_2# check out. **SUMMARY – B3:** #Person_2# finds #Person_2# being mischarged. #Person_1# corrects the bill and #Person_2# pays for it.
UniLMv2: #Person_2# comes to #Person_1#'s annual physical to apply for new health insurance. #Person_1# will do an allergy test, an asthma test, and a blood test.	**UniLMv2:** #Person_2# is checking out. #Person_1# finds #Person_2# has added someone else's laundry service . #Person_1# apologizes and will correct the bill.
Transformer: #Person_2# goes to #Person_1# for an annual physical examination. #Person_1# will send #Person_1# for an asthma test and what #Person_2# eats.	**Transformer:** #Person_2# checks out with #Person_2#'s assistance and thinks they'll be very sorry for the laundry service.

Figure 2: Case study on DialogSum. Dialogue-A - a doctor and a patient dialogue, Dialogue-B - a customer and a hotel service dialogue.

to summarize (Goo and Chen, 2018). Therefore, we conduct error analysis and case studies on DialogSum to quantitatively and qualitatively discuss such challenges.

5.1 Error Analysis

We make error analysis on the 50 selected model-generated summaries (Section 4). Table 11 summarizes the five most frequent error types and their error rates. In general, UniLMv2$_{\text{BASE}}$ shows better performance than Transformer, but its error rates are still high. In particular, incorrect **coreference** (c.f. Section 4) sees the highest error rates for both models, indicating that models can be confused because of interactive information flow. Compared with Transformer, UniLMv2$_{\text{BASE}}$ can greatly avoid errors regarding unfactual information (-52%) and syntactic (-50%). However, it still suffers from coreference issues, and tends to generate redundant summaries.

Error Type	Transformer	UniLMv2$_{\text{BASE}}$
Incorrect Coref.	94%	60%
Missing Salient Inf.	64%	32%
Redundant Inf.	62%	44%
Unfactual Inf.	74%	22%
Syntactic Error	72%	22%

Table 11: Error analysis of model performance on DialogSum. Coref. stands for coreference, and Inf. stands for Information.

5.2 Case Study

We demonstrate two dialogues and their human-annotated/system-generated summaries in Figure 2.

First, a big challenge posed by spoken dialogues is that their **information flow** is different from monologic text, which is intuitively reflected in the dialogue **discourse structures** (Wolf and Gibson, 2005). For example, two utterances can be closely related even where there is a large distance between them. Such phenomena are common in spoken dialogues such as negotiations and procedures (e.g., medical consultation and police reports). Due to the unique structure of the spoken dialogue, important information is rather dispersed than well-structured monologues and written-dialogues.

Regular greetings can be useless to written dialogue summaries (e.g. SAMSum), which is reflected by that LEAD is worse than LONGEST on SAMSum (Table 4). In contrast, LEAD outperforms LONGEST by over 3% on DIALOGSUM. This is because, for spoken dialogues, such utterances sometimes express and indicate essential intents of speakers (c.f. Section 4). **Farewells** also express the dialogue consequence and future plan of the speakers (e.g. dialogues in Figure 2). Besides, **interruptions** appear frequently in the middle of conversations (Figure 2, DIALOGUE-B). These interruptions make other speaker's utterances incomplete, adding redundant information, and can also destroy coherent discourse structures, making dialogues more difficult to encode. These characteristics also make information in DIALOGSUM dialogues more dispersed than existing datasets.

Second, **coreference** and **ellipsis** are frequent in spoken dialogues (Grosz et al., 1995; Quan et al., 2019). It is a natural behavior of communication that humans obey as a rhetorical principle for saving words and avoiding repetitions. Although it can be trivial for humans, their understanding can be challenging to a neural model. For example, to correctly generate "*mischarged/wrong*" in SUMMARY-B1-SUMMARY-B3, models need to understand "*I think you have added someone else's (laundry service on my bill)*", where "*my bill*" refers to "*#Person_2#'s bill*".

Third, **pragmatics** and **social common sense** give a unique challenge for spoken language understanding and has a significant impact on summarization. From the last two sentences of DIALOGUE-B, human could understand that the "*Here you are*" is actually "*make a payment*", and "*Goodbye*" indicates that the event "*check out*" is finished. It requires commonsense knowledge to fully understand such dialogues. Beside, dialogues are summarized from a *different* perspective (compared with speakers' perspective), which suggests that summarizing dialogues needs to go beyond summarizing dialogue contents, but also dialogue actions at the **pragmatic** level. For example, "*explains*" in SUMMARY-A1 and SUMMARY-A2 summarizes multiple dialogue actions of #Person_1#, "*agree*" in Figure 1 (a) summarizes actions of both speakers. It requires model to not only summarize *what speakers are saying*, but also *what they are doing*.

6 Conclusion

We presented DIALOGSUM, a large-scale dialogue summarization dataset, investigating its characteristics and challenges empirically. Experiments with typical models show that DIALOGSUM is highly abstractive, and poses unique challenges in discourse and complex co-references. From these observations, we made discussion on the uniqueness of spoken dialogue summarization, listing several key problems to consider in future modeling. To our knowledge, we are the first to release a large-scale dataset for real-life scenario dialogue summarization.

7 Ethics Consideration

As mentioned, we collect our data from DailyDialog, DREAM and MuTual that all are public for academic use. The additional data are from www.tingroom.com, which are available to the public as well. The sources of our dialogue data are freely accessible online without copyright constraint to academic use.

We hired annotators who have degrees in English Linguistics or Applied Linguistics. Before formal annotation, we annotated 50 samples randomly extracted from the dataset, and calculated our average annotation time so we could set a fair salary for annotators' training annotation. During the training annotation process, they were paid as well. We also calculated the average annotation time for each dialogue during training, based on which we determined the final salary was around 9.5 dollars per hour. This hourly salary was the same for manual checking. All of our annotators took this annotation as a part-time job.

Acknowledgments

Yue Zhang is the corresponding author. We would like to thank the anonymous reviewers for their comments, and Xuefeng Bai, Ming Shen and Bonnie Webber for insightful discussion and proofreading. Also, we would like to thank Leyang Cui for sharing original dialogue data of MuTual dataset. Our sincere appreciation goes to annotators who have contributed to building this dataset. This work receives funding support from the Westlake University and Bright Dream Joint Institute for Intelligent Robotics, and a research grant from Rxhui Inc.

References

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, and Hsiao-Wuen Hon. 2020. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 642–652. PMLR.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

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 the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118, 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.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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.

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.

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.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.

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

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.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. Citeseer.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 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.

Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. 2003. The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461.

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. 2019. Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yang Liu, Sheng Shen, and Mirella Lapata. 2020. Noisy self-knowledge distillation for text summarization. *arXiv preprint arXiv:2009.07032*.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. The Penn Discourse Treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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.

Dragomir R. Radev, Eduard Hovy, and Kathleen McKeown. 2002. Introduction to the special issue on summarization. *Computational Linguistics*, 28(4):399–408.

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

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

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.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

A Dialogue Topics

We use k-means (Likas et al., 2003) to cluster the dialogue topic with GloVe embedding (Pennington et al., 2014), where $k = 20$. Figure 3 presents the proportion of clustering results. Table 12 presents the cluster topics with corresponding id, which is assigned by human.

B Dialogue Turns

The number of conversation turns can have a direct impact on neural models. Multi-turn dialogues correspond to more complicated information flow and discourse structure. Following Chen and Yang (2020), we split test data based on dialogue turns, with a step size of 3, and show model performance on different dialogue turns.

The results are shown in Figure 4. The performance of Transformer and UNILMV2$_{\text{BASE}}$ decreases when number of turns grows, suggesting that more interactions between interlocutors and complicated discourse structures bring challenge. This phenomenon is also observed by (Chen and Yang, 2020) for SAMSum.

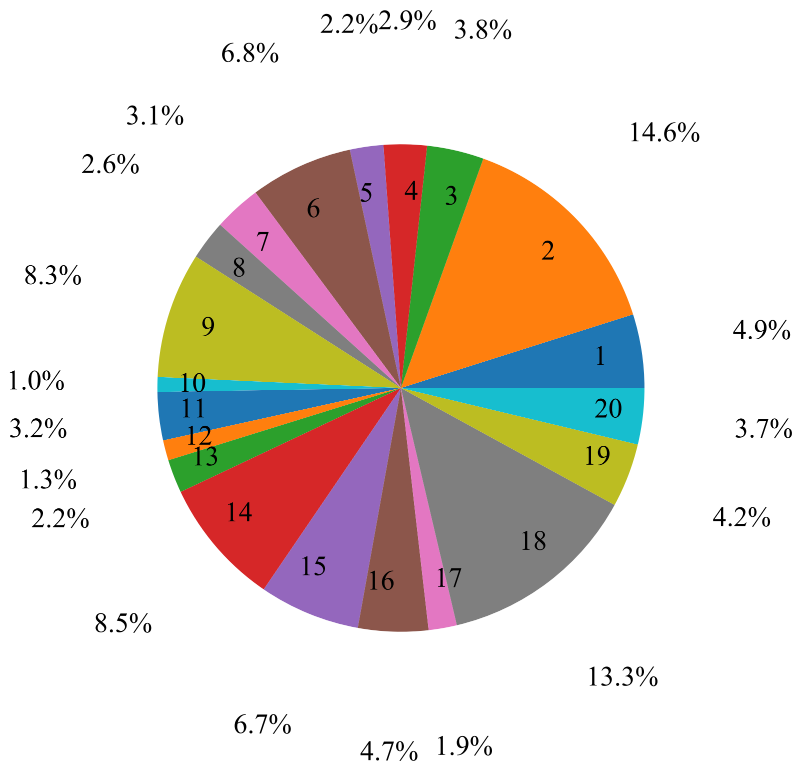

Figure 3: Proportion of dialogue topics.

ID	topic	ID	topic	ID	topic	ID	topic
1	interpersonal relation	6	education	11	personal and business appoinment	16	transportation
2	work and career	7	hobby	12	housing and apartment	17	in-store shopping
3	causal chitchat	8	interview	13	consultation	18	health and medicine
4	hotel and restaurant service	9	vacation plan	14	personal life	19	entertainment
5	sales	10	climate	15	economics	20	food ordering

Table 12: Topic clusters of DIALOGSUM.

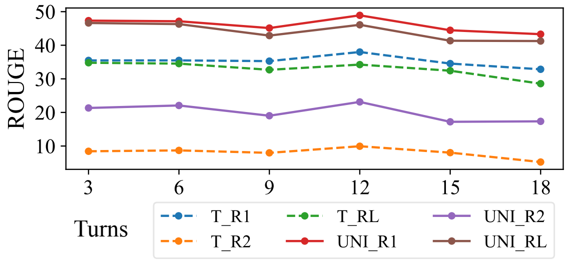

Figure 4: Model performance against the number of dialogue turns. T - Transformer. UNI - UNILMV2.