SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, Aleksander Wawer
Samsung R&D Institute Poland
{b.gliwa, i.mochol, m.biesek, a.wawer}@samsung.com

Abstract

This paper introduces the **SAMSum Corpus**, a new dataset with abstractive dialogue summaries. We investigate the challenges it poses for automated summarization by testing several models and comparing their results with those obtained on a corpus of news articles. We show that model-generated summaries of dialogues achieve higher ROUGE scores than the model-generated summaries of news – in contrast with human evaluators' judgement. This suggests that a challenging task of abstractive dialogue summarization requires dedicated models and non-standard quality measures. To our knowledge, our study is the first attempt to introduce a high-quality chat-dialogues corpus, manually annotated with abstractive summarizations, which can be used by the research community for further studies.

1 Introduction and related work

The goal of the summarization task is condensing a piece of text into a shorter version that covers the main points succinctly. In the abstractive approach important pieces of information are presented using words and phrases not necessarily appearing in the source text. This requires natural language generation techniques with high level of semantic understanding (Chopra et al., 2016; Rush et al., 2015; Khandelwal et al., 2019; Zhang et al., 2019; See et al., 2017; Chen and Bansal, 2018; Gehrmann et al., 2018).

Major research efforts have focused so far on summarization of single-speaker documents like news (e.g., Nallapati et al. (2016)) or scientific publications (e.g., Nikolov et al. (2018)). One of the reasons is the availability of large, high-quality news datasets with annotated summaries, e.g., CNN/Daily Mail (Hermann et al., 2015; Nallapati et al., 2016). Such a comprehensive dataset for dialogues is lacking.

The challenges posed by the abstractive dialogue summarization task have been discussed in the literature with regard to AMI meeting corpus (McCowan et al., 2005), e.g. Banerjee et al. (2015), Mehdad et al. (2014), Goo and Chen (2018). Since the corpus has a low number of summaries (for 141 dialogues), Goo and Chen (2018) proposed to use assigned topic descriptions as gold references. These are short, label-like goals of the meeting, e.g., *costing evaluation of project process; components, materials and energy sources; chitchat*. Such descriptions, however, are very general, lacking the messenger-like structure and any information about the speakers.

To benefit from large news corpora, Ganesh and Dingliwal (2019) built a dialogue summarization model that first converts a conversation into a structured text document and later applies an attention-based pointer network to create an abstractive summary. Their model, trained on structured text documents of CNN/Daily Mail dataset, was evaluated on the Argumentative Dialogue Summary Corpus (Misra et al., 2015), which, however, contains only 45 dialogues.

In the present paper, we further investigate the problem of abstractive dialogue summarization. With the growing popularity of online conversations via applications like Messenger, WhatsApp and WeChat, summarization of chats between a few participants is a new interesting direction of summarization research. For this purpose we have created the **SAMSum Corpus**[1] which contains over 16k chat dialogues with manually annotated summaries. The dataset is freely available for the research community[2].

The paper is structured as follows: in Section 2

[1]The name is a shortcut for **S**amsung **A**bstractive **M**essenger **Sum**marization

[2]The dataset will be published on ELRA language resources catalogue.

Proceedings of the 2nd Workshop on New Frontiers in Summarization, pages 70–79
Hong Kong, China, November 4, 2019.

Dataset	Train	Validation	Test
CNN/DM	287 227	13 368	11 490
SAMSum	14 732	818	819

Table 1: Datasets sizes

we present details about the new corpus and describe how it was created, validated and cleaned. Brief description of baselines used in the summarization task can be found in Section 3. In Section 4, we describe our experimental setup and parameters of models. Both evaluations of summarization models, the automatic with ROUGE metric and the linguistic one, are reported in Section 5 and Section 6, respectively. Examples of models' outputs and some errors they make are described in Section 7. Finally, discussion, conclusions and ideas for further research are presented in sections 8 and 9.

2 SAMSum Corpus

Initial approach. Since there was no available corpus of messenger conversations, we considered two approaches to build it: (1) using existing datasets of documents, which have a form similar to chat conversations, (2) creating such a dataset by linguists.

In the first approach, we reviewed datasets from the following categories: chatbot dialogues, SMS corpora, IRC/chat data, movie dialogues, tweets, comments data (conversations formed by replies to comments), transcription of meetings, written discussions, phone dialogues and daily communication data. Unfortunately, they all differed in some respect from the conversations that are typically written in messenger apps, e.g. they were too technical (IRC data), too long (comments data, transcription of meetings), lacked context (movie dialogues) or they were more of a spoken type, such as a dialogue between a petrol station assistant and a client buying petrol.

As a consequence, we decided to create a chat dialogue dataset by constructing such conversations that would epitomize the style of a messenger app.

Process of building the dataset. Our dialogue summarization dataset contains natural messenger-like conversations created and written down by linguists fluent in English. The style and register of conversations are diversified – dialogues could be informal, semi-formal or formal, they may contain slang phrases, emoticons and typos. We asked linguists to create conversations similar to those they write on a daily basis, reflecting the proportion of topics of their real-life messenger conversations. It includes chit-chats, gossiping about friends, arranging meetings, discussing politics, consulting university assignments with colleagues, etc. Therefore, this dataset does not contain any sensitive data or fragments of other corpora.

Each dialogue was created by one person. After collecting all of the conversations, we asked language experts to annotate them with summaries, assuming that they should (1) be rather short, (2) extract important pieces of information, (3) include names of interlocutors, (4) be written in the third person. Each dialogue contains only one reference summary.

Validation. Since the SAMSum corpus contains dialogues created by linguists, the question arises whether such conversations are really similar to those typically written via messenger apps. To find the answer, we performed a validation task. We asked two linguists to doubly annotate 50 conversations in order to verify whether the dialogues could appear in a messenger app and could be summarized (i.e. a dialogue is not too general or unintelligible) or not (e.g. a dialogue between two people in a shop). The results revealed that 94% of examined dialogues were classified by both annotators as good i.e. they do look like conversations from a messenger app and could be condensed in a reasonable way. In a similar validation task, conducted for the existing dialogue-type datasets (described in the Initial approach section), the annotators agreed that only 28% of the dialogues resembled conversations from a messenger app.

Cleaning data. After preparing the dataset, we conducted a process of cleaning it in a semi-automatic way. Beforehand, we specified a format for written dialogues with summaries: a colon should separate an author of utterance from its content, each utterance is expected to be in a separate line. Therefore, we could easily find all deviations from the agreed structure – some of them could be automatically fixed (e.g. when instead of a colon, someone used a semicolon right after the interlocutor's name at the beginning of an utterance), others were passed for verification to linguists. We also tried to correct typos in interlocutors' names (if one person has several utter-

ances, it happens that, before one of them, there is a typo in his/her name) – we used the Levenshtein distance to find very similar names (possibly with typos e.g. 'George' and 'Goerge') in a single conversation, and those cases with very similar names were passed to linguists for verification.

Description. The created dataset is made of 16369 conversations distributed uniformly into 4 groups based on the number of utterances in conversations: 3-6, 7-12, 13-18 and 19-30. Each utterance contains the name of the speaker. Most conversations consist of dialogues between two interlocutors (about 75% of all conversations), the rest is between three or more people. Table 1 presents the size of the dataset split used in our experiments. The example of a dialogue from this corpus is shown in Table 2.

Dialogue
Blair: Remember we are seeing the wedding planner after work Chuck: Sure, where are we meeting her? Blair: At Nonna Rita's Chuck: Can I order their seafood tagliatelle or are we just having coffee with her? I've been dreaming about it since we went there last month Blair: Haha sure why not Chuck: Well we both remmber the spaghetti pomodoro disaster from our last meeting with Diane Blair: Omg hahaha it was all over her white blouse Chuck: :D Blair: :P
Summary
Blair and Chuck are going to meet the wedding planner after work at Nonna Rita's. The tagliatelle served at Nonna Rita's are very good.

Table 2: Example of a dialogue from the collected corpus

3 Dialogues baselines

The baseline commonly used in the news summarization task is Lead-3 (See et al., 2017), which takes three leading sentences of the document as the summary. The underlying assumption is that the beginning of the article contains the most significant information. Inspired by the Lead-n model, we propose a few different simple models:

Model	**n**	**R-1**	**R-2**	**R-L**
LEAD	3	31.40	8.68	29.42
	4	31.87	8.93	29.91
	5	32.02	9.53	**30.07**
MIDDLE	3	28.04	6.57	26.13
	4	30.08	7.96	28.10
	5	29.91	8.12	27.97
LONGEST	3	**32.46**	10.27	29.92
	4	32.19	**10.35**	29.91
	5	31.61	10.21	29.55
LONGER -THAN	10	28.31	9.69	26.72
	20	29.36	10.23	27.59
	30	29.61	10.28	27.71
MOST-ACTIVE -PERSON	n/a	26.54	8.55	24.57

Table 3: Baselines for the dialogues summarization

- MIDDLE-n, which takes n utterances from the middle of the dialogue,
- LONGEST-n, treating only n longest utterances in order of length as a summary,
- LONGER-THAN-n, taking only utterances longer than n characters in order of length (if there is no such long utterance in the dialogue, takes the longest one),
- MOST-ACTIVE-PERSON, which treats all utterances of the most active person in the dialogue as a summary.

Results of the evaluation of the above models are reported in Table 3. There is no obvious baseline for the task of dialogues summarization. We expected rather low results for Lead-3, as the beginnings of the conversations usually contain greetings, not the main part of the discourse. However, it seems that in our dataset greetings are frequently combined with question-asking or information passing (sometimes they are even omitted) and such a baseline works even better than the MIDDLE baseline (taking utterances from the middle of a dialogue). Nevertheless, the best dialogue baseline turns out to be the LONGEST-3 model.

4 Experimental setup

This section contains a description of setting used in the experiments carried out.

4.1 Data preparation

In order to build a dialogue summarization model, we adopt the following strategies: (1) each candidate architecture is trained and evaluated on the dialogue dataset; (2) each architecture is trained on the train set of CNN/Daily Mail joined together with the train set of the dialogue data, and evaluated on the dialogue test set.

In addition, we prepare a version of dialogue data, in which utterances are separated with a special token called the separator (artificially added token e.g. '<EOU>' for models using word embeddings, '|' for models using subword embeddings). In all our experiments, news and dialogues are truncated to 400 tokens, and summaries – to 100 tokens. The maximum length of generated summaries was not limited.

4.2 Models

We carry out experiments with the following summarization models (for all architectures we set the beam size for beam search decoding to 5):

- **Pointer generator network** (See et al., 2017). In the case of *Pointer Generator*, we use a default configuration[3], changing only the minimum length of the generated summary from 35 (used in news) to 15 (used in dialogues).
- **Transformer** (Vaswani et al., 2017). The model is trained using OpenNMT library[4]. We use the same parameters for training both on news and on dialogues[5], changing only the minimum length of the generated summary – 35 for news and 15 for dialogues.
- **Fast Abs RL** (Chen and Bansal, 2018). It is trained using its default parameters[6]. For dialogues, we change the convolutional word-level sentence encoder (used in extractor part) to only use kernel with size equal 3 instead of 3-5 range. It is caused by the fact that some of utterances are very short and the default setting is unable to handle that.
- **Fast Abs RL Enhanced**. The additional variant of the *Fast Abs RL* model with slightly changed utterances i.e. to each utterance, at the end, after artificial separator, we add names of all other interlocutors. The reason for that is that *Fast Abs RL* requires text to be split into sentences (as it selects sentences and then paraphrase each of them). For dialogues, we divide text into utterances (which is a natural unit in conversations), so sometimes, a single utterance may contain more than one sentence. Taking into account how this model works, it may happen that it selects an utterance of a single person (each utterance starts with the name of the author of the utterance) and has no information about other interlocutors (if names of other interlocutors do not appear in selected utterances), so it may have no chance to use the right people's names in generated summaries.
- **LightConv and DynamicConv** (Wu et al., 2019). The implementation is available in fairseq[7] (Ott et al., 2019). We train lightweight convolution models in two manners: (1) learning token representations from scratch; in this case we apply BPE tokenization with the vocabulary of 30K types, using fastBPE implementation[8] (Sennrich et al., 2015); (2) initializing token embeddings with pre-trained language model representations; as a language model we choose GPT-2 small (Radford et al., 2019).

4.3 Evaluation metrics

We evaluate models with the standard ROUGE metric (Lin, 2004), reporting the F_1 scores (with stemming) for ROUGE-1, ROUGE-2 and ROUGE-L following previous works (Chen and Bansal, 2018; See et al., 2017). We obtain scores using the `py-rouge` package[9].

5 Results

The results for the news summarization task are shown in Table 4 and for the dialogue summarization – in Table 5. In both domains, the best models' ROUGE-1 exceeds 39, ROUGE-2 – 17 and

[3] https://github.com/abisee/pointer-generator
[4] https://github.com/OpenNMT/OpenNMT-py
[5] http://opennmt.net/OpenNMT-py/Summarization.html
[6] https://github.com/ChenRocks/fast_abs_rl
[7] https://github.com/pytorch/fairseq
[8] https://github.com/glample/fastBPE
[9] https://pypi.org/project/py-rouge/

ROUGE-L – 36. Note that the strong baseline for news (Lead-3) is outperformed in all three metrics only by one model. In the case of dialogues, all tested models perform better than the baseline (LONGEST-3).

In general, the Transformer-based architectures benefit from training on the joint dataset: news+dialogues, even though the news and the dialogue documents have very different structures. Interestingly, this does not seem to be the case for the *Pointer Generator* or *Fast Abs RL* model.

The inclusion of a separation token between dialogue utterances is advantageous for most models – presumably because it improves the discourse structure. The improvement is most visible when training is performed on the joint dataset.

Having compared two variants of the *Fast Abs RL* model – with original utterances and with enhanced ones (see Section 4.2), we conclude that enhancing utterances with information about the other interlocutors helps achieve higher ROUGE values.

The largest improvement of the model performance is observed for *LightConv* and *DynamicConv* models when they are complemented with pretrained embeddings from the language model *GPT-2*, trained on enormous corpora.

It is also worth noting that some models (*Pointer Generator*, *Fast Abs RL*), trained only on the dialogues corpus (which has 16k dialogues), reach similar level (or better) in terms of ROUGE metrics than models trained on the CNN/DM news dataset (which has more than 300k articles). Adding pretrained embeddings and training on the joined dataset helps in achieving significantly higher values of ROUGE for dialogues than the best models achieve on the CNN/DM news dataset.

According to ROUGE metrics, the best performing model is *DynamicConv* with *GPT-2* embeddings, trained on joined news and dialogue data with an utterance separation token.

6 Linguistic verification of summaries

ROUGE is a standard way of evaluating the quality of machine generated summaries by comparing them with reference ones. The metric based on n-gram overlapping, however, may not be very informative for abstractive summarization, where paraphrasing is a keypoint in producing high-quality sentences. To quantify this conjecture, we manually evaluated summaries generated by the models for 150 news and 100 dialogues. We asked two linguists to mark the quality of every summary on the scale of -1, 0, 1, where -1 means that a summarization is poor, extracts irrelevant information or does not make sense at all, 1 – it is understandable and gives a brief overview of the text, and 0 stands for a summarization that extracts only a part of relevant information, or makes some mistakes in the produced summary.

Model	**R-1**	**R-2**	**R-L**
Lead-3 baseline	40.24	17.44	34.90
Pointer Generator	38.72	16.67	35.59
Fast Abs RL	**40.99**	**17.72**	**38.30**
Transformer	38.72	16.89	35.74
LightConv	39.44	17.20	36.20
DynamicConv	39.46	17.33	36.29
LightConv + GPT2 emb	39.52	17.31	36.15
DynamicConv + GPT2 emb	39.94	17.56	36.51

Table 4: Model evaluation on the news corpus test set

We noticed a few annotations (7 for news and 4 for dialogues) with opposite marks (i.e. one annotator judgement was -1, whereas the second one was 1) and decided to have them annotated once again by another annotator who had to resolve conflicts. For the rest, we calculated the linear weighted Cohen's kappa coefficient (McHugh, 2012) between annotators' scores. For news examples, we obtained agreement on the level of 0.371 and for dialogues – 0.506. The annotators' agreement is higher on dialogues than on news, probably because of structures of those data – articles are often long and it is difficult to decide what the key-point of the text is; dialogues, on the contrary, are rather short and focused mainly on one topic.

For manually evaluated samples, we calculated ROUGE metrics and the mean of two human ratings; the prepared statistics is presented in Table 6. As we can see, models generating dialogue summaries can obtain high ROUGE results, but their outputs are marked as poor by human annotators. Our conclusion is that the ROUGE metric corresponds with the quality of generated summaries for news much better than for dialogues, confirmed by Pearson's correlation between human evaluation and the ROUGE metric, shown

Model	Train data	Separator	R-1	R-2	R-L
LONGEST-3 baseline	n/a	n/a	32.46	10.27	29.92
Pointer Generator	dialogues	no	38.55	14.14	34.85
Pointer Generator	dialogues	yes	40.08	15.28	36.63
Fast Abs RL	dialogues	no	40.96	17.18	39.05
Fast Abs RL Enhanced	dialogues	no	41.95	18.06	39.23
Transformer	dialogues	no	36.62	11.18	33.06
Transformer	dialogues	yes	37.27	10.76	32.73
LightConv	dialogues	no	33.19	11.14	30.34
DynamicConv	dialogues	no	33.79	11.19	30.41
DynamicConv	dialogues	yes	33.69	10.88	30.93
LightConv + GPT-2 emb.	dialogues	no	41.81	16.34	37.63
DynamicConv + GPT-2 emb.	dialogues	no	41.79	16.44	37.54
DynamicConv + GPT-2 emb.	dialogues	yes	41.54	16.29	37.07
Pointer Generator	news + dialogues	no	35.04	13.25	32.42
Pointer Generator	news + dialogues	yes	37.27	14.42	34.36
Fast Abs RL	news + dialogues	no	41.03	16.93	39.05
Fast Abs RL Enhanced	news + dialogues	no	41.87	17.47	39.53
Transformer	news + dialogues	no	41.91	18.25	38.77
Transformer	news + dialogues	yes	42.37	18.44	39.27
LightConv	news + dialogues	no	40.29	17.28	36.81
DynamicConv	news + dialogues	no	40.66	17.41	37.20
DynamicConv	news + dialogues	yes	41.07	17.11	37.27
LightConv + GPT-2 emb.	news + dialogues	no	44.47	19.75	40.07
DynamicConv + GPT-2 emb.	news + dialogues	no	44.69	20.28	40.76
DynamicConv + GPT-2 emb.	news + dialogues	yes	**45.41**	**20.65**	**41.45**

Table 5: Model evaluation on the dialogues corpus test set

in Table 7.

7 Difficulties in dialogue summarization

In a structured text, such as a news article, the information flow is very clear. However, in a dialogue, which contains discussions (e.g. when people try to agree on a date of a meeting), questions (one person asks about something and the answer may appear a few utterances later) and greetings, most important pieces of information are scattered across the utterances of different speakers. What is more, articles are written in the third-person point of view, but in a chat everyone talks about themselves, using a variety of pronouns, which further complicates the structure. Additionally, people talking on messengers often are in a hurry, so they shorten words, use the slang phrases (e.g. 'u r gr8' means 'you are great') and make typos. These phenomena increase the difficulty of performing dialogue summarization.

Table 8 and 9 show a few selected dialogues, together with summaries produced by the best tested models:

- *DynamicConv + GPT-2* embeddings with a separator (trained on news + dialogues),
- *DynamicConv + GPT-2* embeddings (trained on news + dialogues),
- *Fast Abs RL* (trained on dialogues),
- *Fast Abs RL Enhanced* (trained on dialogues),
- *Transformer* (trained on news + dialogues).

One can easily notice problematic issues. Firstly, the models frequently have difficulties in associating names with actions, often repeating the same name, e.g., for Dialogue 1 in Table 8, *Fast Abs RL* generates the following summary: 'lilly and lilly are going to eat salmon'. To help the model deal with names, the utterances are enhanced by adding information about the other interlocutors – *Fast Abs RL enhanced* variant de-

		#examples	mean	median	R-1	R-2	R-L
NEWS	overall	100	0.18	0.5	39.76	16.55	36.23
	Fast Abs RL	50	0.33	0.5	42.33	18.28	38.82
	DynamicConv	50	0.03	0.25	37.19	14.81	33.64
DIALOGUES	overall	150	-0.503	-0.5	43.53	19.94	40.66
	Fast Abs RL	50	-0.55	-0.75	42.16	19.28	40.37
	Fast Abs RL Enhanced	50	-0.63	-1.0	39.79	16.59	37.05
	DynamicConv + GPT-2 emb.	50	-0.33	-0.5	48.63	23.95	44.57

Table 6: Statistics of human evaluation of summaries' quality and ROUGE evaluation of those summaries

	ROUGE-1		ROUGE-2		ROUGE-L	
	corr	p-value	corr	p-value	corr	p-value
NEWS	0.47	1e-6	0.44	6e-6	0.48	1e-6
DIALOGUES	0.32	7.7e-5	0.30	1.84e-4	0.32	8.1e-5

Table 7: Pearson's correlations between human judgement and ROUGE metric

scribed in Section 4.2. In this case, after enhancement, the model generates a summary containing both interlocutors' names: 'lily and gabriel are going to pasta...'. Sometimes models correctly choose speakers' names when generating a summary, but make a mistake in deciding who performs the action (the subject) and who receives the action (the object), e.g. for Dialogue 4 *DynamicConv + GPT-2 emb. w/o sep.* model generates the summary 'randolph will buy some earplugs for maya', while the correct form is 'maya will buy some earplugs for randolph'.

A closely related problem is capturing the context and extracting information about the arrangements after the discussion. For instance, for Dialogue 4, the *Fast Abs RL* model draws a wrong conclusion from the agreed arrangement. This issue is quite frequently visible in summaries generated by *Fast Abs RL*, which may be the consequence of the way it is constructed; it first chooses important utterances, and then summarizes each of them separately. This leads to the narrowing of the context and loosing important pieces of information.

One more aspect of summary generation is deciding which information in the dialogue content is important. For instance, for Dialogue 3 *DynamicConv + GPT-2 emb. with sep.* generates a correct summary, but focuses on a piece of information different than the one included in the reference summary. In contrast, some other models – like *Fast Abs RL enhanced* – select both of the pieces of information appearing in the discussion. On the other hand, when summarizing Dialogue 5, the models seem to focus too much on the phrase 'it's the best place', intuitively not the most important one to summarize.

8 Discussion

This paper is a step towards abstractive summarization of dialogues by (1) introducing a new dataset, created for this task, (2) comparison with news summarization by the means of automated (ROUGE) and human evaluation.

Most of the tools and the metrics measuring the quality of text summarization have been developed for a single-speaker document, such as news; as such, they are not necessarily the best choice for conversations with several speakers.

We test a few general-purpose summarization models. In terms of human evaluation, the results of dialogues summarization are worse than the results of news summarization. This is connected with the fact that the dialogue structure is more complex – information is spread in multiple utterances, discussions, questions, more typos and slang words appear there, posing new challenges for summarization. On the other hand, dialogues are divided into utterances, and for each utterance its author is assigned. We demonstrate in experiments that the models benefit from the introduction of separators, which mark utterances for each person. This suggests that dedicated models having some architectural changes, taking into ac-

Dialogue 1	**Dialogue 2**
1. lilly: sorry, i'm gonna be late 2. lilly: don't wait for me and order the food 3. gabriel: no problem, shall we also order something for you? 4. gabriel: so that you get it as soon as you get to us? 5. lilly: good idea 6. lilly: pasta with salmon and basil is always very tasty here	1. randolph: honey 2. randolph: are you still in the pharmacy? 3. maya: yes 4. randolph: buy me some earplugs please 5. maya: how many pairs? 6. randolph: 4 or 5 packs 7. maya: i'll get you 5 8. randolph: thanks darling
REF: lilly will be late. gabriel will order pasta with salmon and basil for her. **L3:** 6, 3, 4 [38/17/38] **DS:** lilly and gabriel are going to order pasta with salmon and basil [62/42/62] **D:** lilly and gabriel are going to order pasta with salmon and basil [62/42/62] **F:** lilly will be late . she will order the food . lilly and lilly are going to eat salmon and basil [55/39/55] **FE:** lilly will be late . lilly and gabriel are going to pasta with salmon and basil is always tasty . [63/47/63] **T:** lilly will order the food as soon as she gets to gabriel [31/17/23]	**REF:** maya will buy 5 packs of earplugs for randolph at the pharmacy. **L3:** 2, 4, 8 [36/8/36] **DS:** randolph and maya are going to buy some earplugs for randolph. [43/19/43] **D:** randolph will buy some earplugs for maya. [63/24/42] **F:** maya is in the pharmacy . maya will get 5 . [48/21/48] **FE**: randolph is in the pharmacy . randolph will buy some earplugs for randolph . maya will get 5 . [64/38/64] **T:** randolph will buy some earplugs for randolph . maya will get 5 pairs . [58/36/42]

Table 8: Examples of dialogues (Part 1). REF – reference summary, L3 – LONGEST-3 baseline, DS – DynamicConv + GPT-2 emb. with sep., D – DynamicConv + GPT-2 emb., F – Fast Abs RL, FE – Fast Abs RL Enhanced, T – Transformer. For L3, three longest utterances are listed. Rounded ROUGE values [R-1/R-2/R-L] are given in square brackets.

count the assignation of a person to an utterance in a systematic manner, could improve the quality of dialogue summarization.

We show that the most popular summarization metric ROUGE does not reflect the quality of a summary. Looking at the ROUGE scores, one concludes that the dialogue summarization models perform better than the ones for news summarization. In fact, this hypothesis is not true – we performed an independent, manual analysis of summaries and we demonstrated that high ROUGE results, obtained for automatically-generated dialogue summaries, correspond with lower evaluation marks given by human annotators. An interesting example of the misleading behavior of the ROUGE metrics is presented in Table 9 for Dialogue 4, where a wrong summary – 'paul and cindy don't like red roses.' – obtained all ROUGE values higher than a correct summary – 'paul asks cindy what color flowers should buy.'. Despite lower ROUGE values, news summaries were scored higher by human evaluators. We conclude that when measuring the quality of model-generated summaries, the ROUGE metrics are more indicative for news than for dialogues, and a new metric should be designed to measure the quality of abstractive dialogue summaries.

9 Conclusions

In our paper we have studied the challenges of abstractive dialogue summarization. We have addressed a major factor that prevents researchers from engaging into this problem: the lack of a proper dataset. To the best of our knowledge, this is the first attempt to create a comprehensive resource of this type which can be used in future research. The next step could be creating an even more challenging dataset with longer dialogues that not only cover one topic, but span over

Dialogue 3	**Dialogue 4**
1. ashleigh: looks like we're going to the cinema!! 2. ashleigh: <file_gif> 3. peter: you got the job?? 4. ashleigh: i got hte job! :d 5. peter: <file_gif> 6. ashleigh: <file_gif>	1. paul: what color flowers should i get 2. cindy: any just not yellow 3. paul: ok, pink? 4. cindy: no maybe red 5. paul: just tell me what color and what type ok? 6. cindy: ugh, red roses!
REF: ashleigh got the job.	**REF:** paul will buy red roses following cindy's advice.
L3: 1, 4, 3 [33/18/33] **DS:** ashleigh and peter are going to the cinema. [33/0/33] **D:** ashleigh got hte job. [75/33/75]	**L3:** 5, 1, 2 [13/0/13] **DS:** paul and cindy don't like red roses. [47/13/35] **D:** paul asks cindy what color flowers should buy. [35/0/24]
F: ashleigh and ashleigh are going to the cinema. peter got the job . [50/29/50] **FE:** ashley and peter are going to the cinema together . ashleigh got the job . [47/40/47] **T:** ashleigh got the job at the cinema . peter and ashleigh are going there . [47/40/47]	**F:** cindy is going to buy red roses [50/29/38] **FE:** cindy is buying red roses . cindy will buy red . [56/38/44] **T:** cindy does n't know what color should get. cindy does not know what to do [8/0/8]

Dialogue 5
1. eve: where are we meeting? 2. charlie: at the entrance 3. nicole: yes, it's the best place. we would't find each other inside, it'll be too crowded 4. eve: ok!
REF: eve, charlie and nicole are meeting at the entrance.
L3: 3, 1, 2 [43/11/43] **DS:** eve, charlie and nicole are meeting at the entrance. [100/100/100] **D:** eve, charlie and nicole are meeting at the entrance. [100/100/100] **F:** charlie is at the entrance . it 's the best place . [42/24/42] **FE:** charlie is at the entrance . nicole and charlie are going to find each other inside . [58/18/42] **T:** eve and nicole are meeting at the entrance . it 's the best place to meet . [67/55/67]

Table 9: Examples of dialogues (Part 2). REF – reference summary, L3 – LONGEST-3 baseline, DS – DynamicConv + GPT-2 emb. with sep., D – DynamicConv + GPT-2 emb., F – Fast Abs RL, FE – Fast Abs RL Enhanced, T – Transformer. For L3, three longest utterances are listed. Rounded ROUGE values [R-1/R-2/R-L] are given in square brackets.

numerous different ones.

As shown, summarization of dialogues is much more challenging than of news. In order to perform well, it may require designing dedicated tools, but also new, non-standard measures to capture the quality of abstractive dialogue summaries in a relevant way. We hope to tackle these issues in future work.

Acknowledgments

We would like to express our sincere thanks to Tunia Błachno, Oliwia Ebebenge, Monika Jędras and Małgorzata Krawentek for their huge contribution to the corpus collection – without their ideas, management of the linguistic task and verification of examples we would not be able to create this paper. We are also grateful for the reviewers' helpful comments and suggestions.

References

Siddhartha Banerjee, Prasenjit Mitra, and Kazunari Sugiyama. 2015. Abstractive meeting summarization using dependency graph fusion. In *Proceedings of the 24th International Conference on World Wide Web*, pages 5–6.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 675–686.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–âĂŞ98.

Prakhar Ganesh and Saket Dingliwal. 2019. Abstractive summarization of spoken and written conversation. *arXiv:1902.01615*.

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.

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

Karl M. Hermann, TomÃąs KociskÃ¡, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *CoRR*, abs/1506.03340.

Urvashi Khandelwal, Kevin Clark, Dan Jurafsky, and Lukasz Kaiser. 2019. Sample efficient text summarization using a single pre-trained transformer. *CoRR*, abs/1905.08836.

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.

I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, Dennis Reidsma, and P. Wellner. 2005. The ami meeting corpus. In *Proceedings of Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research*, pages 137–140.

Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Yashar Mehdad, Giuseppe Carenini, and Raymond T. Ng. 2014. Abstractive summarization of spoken and written conversations based on phrasal queries. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1220–1230.

Amita Misra, Pranav Anand, Jean Fox Tree, and Marilyn Walker. 2015. Using summarization to discover argument facets in online idealogical dialog. In *The North American Chapter of the Association for Computational Linguistics (NAACL)*.

Ramesh Nallapati, Bowen Zhou, Cicero Nogueira dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Computational Natural Language Learning*.

Nikola Nikolov, Michael Pfeiffer, and Richard Hahnloser. 2018. Data-driven summarization of scientific articles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1073–1083.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

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*, pages 5998–6008.

Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. In *International Conference on Learning Representations*.

Haoyu Zhang, Jianjun Xu, and Ji Wang. 2019. Pretraining-based natural language generation for text summarization. *CoRR*, abs/1902.09243.