ForumSum: A Multi-Speaker Conversation Summarization Dataset

Misha Khalman, Yao Zhao, Mohammad Saleh
Google Research, Brain Team
{khalman,yaozhaoyz,msaleh}@google.com

Abstract

Abstractive summarization quality had large improvements since recent language pretraining techniques. However, currently there is a lack of datasets for the growing needs of conversation summarization applications. Thus we collected ForumSum[1], a diverse and high-quality conversation summarization dataset with human written summaries. The conversations in ForumSum dataset are collected from a wide variety of internet forums. To make the dataset easily expandable, we also release the process of dataset creation. Our experiments show that models trained on ForumSum have better zero-shot and few-shot transferability to other datasets than the existing large chat summarization dataset SAMSum. We also show that using a conversational corpus for pre-training improves the quality of the chat summarization model.

1 Introduction

With increasing number of digital communications, there is an increasing need to manage the exploding amount of information. One way of relieving users from information overload in chat applications is through automatic abstractive summarization by selecting important pieces of information and writing them into accurate, fluent and concise summaries.

Recently there has been a lot of advances in automatic abstractive summarization using large pretrained language models (Zhang et al., 2020; Lewis et al., 2020; Raffel et al., 2020) and finetuning them on downstream summarization datasets. However, most of the pre-training and finetuning domains are news documents (Narayan et al., 2018; See et al., 2017) and there is a lack of attention to summarizing conversations.

In this work, we aim to build a high quality conversation summarization system that generalizes well by creating a new dataset and improving pre-training methods for conversation summarization.

Our contributions include:

- We collected a diverse and high-quality conversational summarization dataset from 281 internet forums and release the dataset creation process to make it easily expandable.
- Our experiments show that models trained on ForumSum transfer better to new domains compared to SAMSum dataset.
- We show that pre-training on conversational corpus improves the quality of chat summarization models.

2 Related Works

SAMSum (Gliwa et al., 2019) is a dataset of 16k high-quality chat-dialogues corpus and their abstractive dialogue summaries manually written by linguists. Linguists are asked to create informal, semi-formal and formal conversations similar to their daily messenger conversations including chit-chats, gossiping about friends, arranging meetings, discussing politics, consulting university assignments with colleagues, etc. Despite its large size and excellent quality, the conversations styles are relatively homogeneous and 75% of the conversations are between two people, whereas summarizing conversations that involve many speakers is a more useful scenario in real world applications.

Ubuntu/NYC (Bhatia et al., 2014) is an online thread summarization datasets that contains 100 threads from ubuntuforums.org and tripadvisor.com and their human written summaries.

BC3 (Ulrich et al., 2008) consists of 40 email threads each annotated with three summaries by three different annotators. Each summary sentence is also annotated with references to the corresponding lines in the emails.

[1] The dataset is available at tensorflow dataset and huggingface.

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4592–4599
November 7–11, 2021. ©2021 Association for Computational Linguistics

We employ BC3, Ubuntu and NYC datasets to evaluate transferability of models trained ForumSum and SAMSum datasets.

MediaSum, SummScreen MediaSum (Zhu et al., 2021) and SummScreen (Chen et al., 2021) are conversations summarization datasets that use speech transcripts as input and automatically mine summaries from interview overviews and TV shows recaps. While being very large and diverse they contain automatically mined summaries which might suffer from lower quality.

AMI, ICSI AMI (McCowan et al., 2005) and ICSI (Shriberg et al., 2004) are meetings transcripts datasets annotated with abstractive summaries.

Meeting transcripts are much longer than messenger or email threads. For example, an average AMI transcript contains 289 turns, while the average number of turns in ForumSum dataset is around 10. Meetings transcripts contain more repetitions, backchannel responses and interjections.

In this paper we focus on summarizing online messaging conversations. Therefore we do not use MediaSum, SummScreen, AMI or ICSI for transferability studies.

3 ForumSum Dataset

Motivated by the lack of diverse multi-speaker conversation summarization datasets we collected ForumSum: a conversation summarization dataset from internet forums labeled with human-written summaries.

First we collected a list of message board websites. We only kept websites that are relatively popular, and are not present in a blocklist of not appropriate websites. The result contains 281 websites.

3.1 Conversation Selection

We scraped all posts with comments from the forums. We combine topic starters and corresponding sequences of comments into conversations.

To get a cleaner and more diverse set of conversations we applied the following filters:

- Filtered out conversations that contained scraping artifacts such as XML tags

- Filtered out conversations that contain any offensive word from a list of of English obscene words and collocations. [2]

- Sample 200 maximum conversations per website to smooth the websites distribution.

- Filtered out conversations where there is only a single speaker.

- Filtered out short conversations that has less than 4 turns.

Conversations that passed this set of filters were sent to be annotated with summaries.

3.2 Crowd-source Annotation

We used Amazon Mechanical Turk to annotate the conversations with human-written summaries.

To guard summary quality we split the dataset into batches of 100-200 examples and sent the conversations to annotators batch-by-batch. After acquiring the results of each batch we manually assessed the summaries quality on a scale from 1 to 5, assessed common issues and made changes to the instructions.

After 4 such iterations we stopped making changes into the instructions set. However, we kept evaluating samples of summaries in each batch to ensure the quality of the summaries does not drop. All batches after the finalized instructions got uniformly good average scores between 4.3 and 4.7. Only batches written after finalized instructions are included in the ForumSum dataset.

See Appendix B for full final instructions.

3.3 ForumSum Style and Format

ForumSum conversations are formatted similarly as conversations in the SAMSum dataset: each utterance starts on a new line, contains an author name and a message text that separated with a colon.

ForumSum summaries also has similar third person style as SAMSum summaries, but are longer and more descriptive. See Appendix A for examples of ForumSum conversations and summaries.

3.4 Statistics

Table 1 and Figure 1 show that ForumSum dataset contains significantly more multi-speaker threads, longer utterances than SAMSum and their distributions are more spread out. ForumSum summaries are longer than SAMSum summaries on average.

[2] https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words/blob/master/en

Dataset	ForumSum	SAMSum
Examples	4058	16369
Input Avg # words	303.45	93.79
Input Avg # turns	10.13	11.17
Input Avg # speakers	6.73	2.40
Target Avg # words	35.95	20.30
Input total # words	1.18M	1.34M
Input total # unique words	55.1k	33.1k

Table 1: ForumSum vs SAMSum dataset statistics.

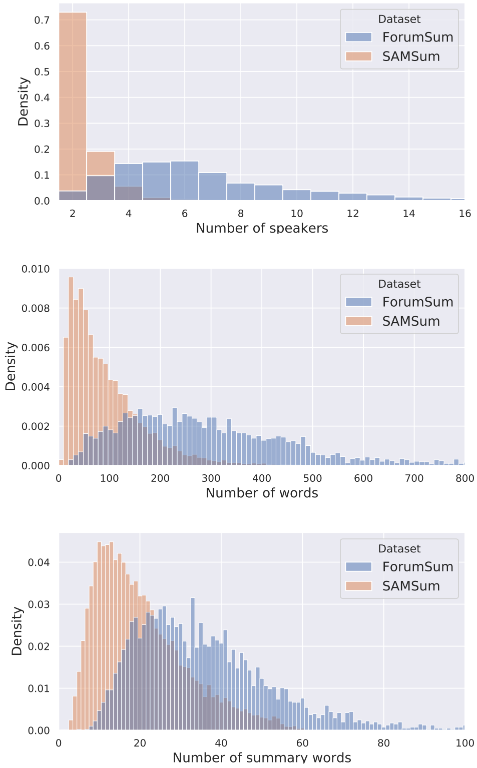

Figure 1: ForumSum vs SAMSum length statistics.

Also, ForumSum conversations have richer vocabulary: they contain 66% more unique words for approximately same number of total words.

4 Experiments and Results

Following (Zhang et al., 2020), we conduct our studies using transformer encoder-decoder models at the BASE size. It had $L = 12, H = 768, F = 3072, A = 12$ where L denotes the number of layers for encoder and decoder (i.e. Transformer blocks), H for the hidden size, F for the feed-forward layer size and A for the number of self-attention heads.

4.1 Zero-shot and Few-shot Transferability

We finetune a pretrained Pegasus-Base (Zhang et al., 2020) model on ForumSum and SAMSum datasets respectively and then study their transferability to out-of-domain chat summarization datasets. We chose to evaluate on Ubuntu/NYC/BC3 because they are very small and contain online-messaging conversations. None of them overlaps with SAMSum/ForumSum datasets.

In zero-shot setting, we directly evaluate models' performance and in few-shot settings we finetune on all training examples in those small datasets.

For BC3 dataset we treat each original summary sentence as an independent summary and construct synthetic conversations using annotated references to email lines. Then we format all input data in Ubuntu/NYC/BC3 consistently with SAMSum/ForumSum as described in Section 3.3.

Table 2 show that finetuning Pegasus first on either SAMSum or ForumSum and then to other smaller datasets improves models performance. Furthermore, ForumSum models transfer to Ubuntu/NYC/BC3 better than SAMSum models in both zero-shot and few-shot settings. This all suggests the variety of conversation distribution in ForumSum help generalization to out of domain datasets. More experimental details are found in Appendix D.

4.2 Human Evaluation

We conducted side-by-side human evaluation comparing the predictions from Pegasus+SAMSum and Pegasus+ForumSum on the test sets of SAMSum, ForumSum, Ubuntu and NYC. Trained human raters, given a chat thread of two summaries in randomized order, are asked to rater compare them in seven categories. More details can be found in Appendix E.

Table 3 show the distribution of human rater's preferences on all downstream domains. SAMSum and ForumSum model both perform better when evaluated on the domains they are trained on. ForumSum models generalize better to other domains such as Ubuntu. Those findings are aligned with the ROUGE scores in Section 4.1.

4.3 Dataset Expansion

Can further dataset expansion potentially improve the quality of our models? To answer that question we evaluated models trained on different number of examples randomly chosen from the training dataset. Extrapolating the relation between quality and number of training examples, we can further predict if adding more data from the same distribution would lead to quality improvements.

initial checkpoint	transfer style	Ubuntu	NYC	BC3
		$R_1/R_2/R_L$	$R_1/R_2/R_L$	$R_1/R_2/R_L$
Pegasus	zero-shot	**32.01 / 18.02/26.51**	**26.87**/7.652/17.6	26.25/9.04/22.47
Pegasus+SAMSum	zero-shot	23.91/9.606/18.7	24.96/9.457/18.54	27.77/9.31/23.77
Pegasus+ForumSum	zero-shot	**31.29**/14.24/24.96	**26.53/11.52/19.83**	**30.53/12.44/26.44**
Pegasus	finetune	50.04/32.18/43.57	33.56/15.02/25.34	32.01/12.79/27.80
Pegasus+SAMSum	finetune	50.5/30.9/42.52	**41.84/20.01/30.4**	34.67/13.46/30.20
Pegasus+ForumSum	finetune	**54.32/36.79/47.43**	**41.61/20.52/31.09**	**36.67/16.59/32.22**

Table 2: Zero-shot and few-shot transferability to smaller datasets. Pegasus+SAMSum/Pegasus+ForumSum refer to a Pegasus pretrained model finetuned on SAMSum/ForumSum respectively. Best numbers within confidence intervals are in bold.

	Evaluation Domain			
	SAMSum	ForumSum	NYC	Ubuntu
ForumSum much worse	18.1%	8.6%	12.1%	4.5%
ForumSum worse	15.3%	8.5%	15.2%	9.3%
ForumSum slightly worse	7.4%	6.6%	8.9%	6.5%
same	34.5%	24.7%	26.6%	37.1%
ForumSum slightly better	7.3%	8.5%	10.3%	11.0%
ForumSum better	13.9%	15.1%	15.6%	13.1%
ForumSum much better	13.4%	18.3%	11.3%	18.6%
overall score	-0.17	0.44	0.0	0.54

Table 3: Side-by-side human evaluation comparing models trained on ForumSum and SAMSum datasets and evaluated on domains without any finetuning.

As shown in Figure 2, both datasets benefit from more training examples as ROUGE scores go up, suggesting further dataset expansion for both SAMSum and ForumSum datasets would further improve summary quality. More details in Appendix D.

4.4 Effect of Pretraining Corpus

We studied whether pretraining on conversational data helps conversation summarization models.

We collected Forums corpus in a similar way we collected source data for ForumSum dataset. To make the pre-training corpus as large as possible we used 56569 forums and didn't apply any filters described in 3.1, but removed all examples included in the ForumSum dataset from the corpus. The pre-training corpus contained around 516M conversations.

We pretrained Pegasus-Base model on the forum corpus for 600K steps and finetuned them on SAMSum and ForumSum datasets.

As shown in Table 4, pretraining on conversational corpora improves conversation summarization models' performance. See Appendix D for details and experiments hyper-parameters.

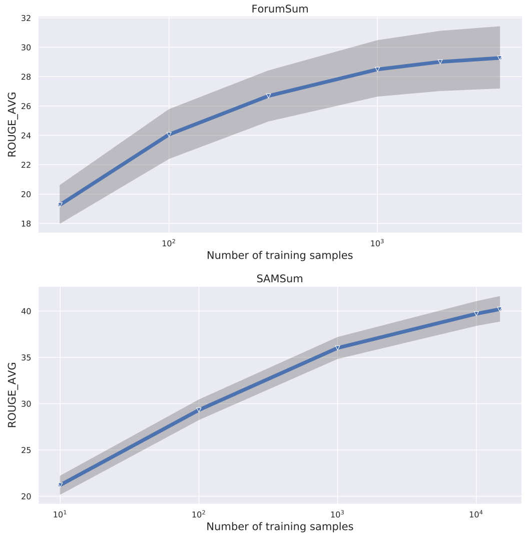

Figure 2: Number of training examples vs the average of ROUGE 1/2/L score.

Pretraining corpus	finetune style	Evaluation	
		SAMSum	ForumSum
		$R_1/R_2/R_L$	$R_1/R_2/R_L$
web/news	full	50.78/26.96/42.85	38.94/16.56/32.31
forums	full	**52.13/28.15/43.99**	**39.71/18.16/32.85**
web/news	few-shot	40.03/15.66/32.26	33.02/11.86/27.29
forums	few-shot	**43.67/19.21/35.73**	**34.56/13.19/27.71**

Table 4: Comparing models pretrained on different corpus and evaluated on conversation summarization tasks. For few-shot experiments we trained on 100 examples.

5 Conclusions

We collected ForumSum, a diverse and high-quality chat summarization dataset with human written summaries. ForumSum can be easily expanded to further improve conversation summarization quality using the released process of dataset creation. Our experiments show that models trained on ForumSum have good zero-shot and few-shot transferability to other conversation summarization datasets measured by ROUGE scores and human evaluations. We also show that using a conversational corpus for pre-training improves the quality of the conversation summarization model.

Acknowledgements

We thank Yinan Wang for helping with human evaluation and Peter J.Liu for valuable feedback on the paper.

References

Sumit Bhatia, Prakhar Biyani, and Prasenjit Mitra. 2014. Summarizing online forum discussions–can dialog acts of individual messages help? In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 2127–2131.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2021. Summscreen: A dataset for abstractive screenplay summarization. *arXiv preprint arXiv:2104.07091*.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *Proceedings of the 2nd Workshop on New Frontiers in Summarization*.

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.

Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, M Guillemot, Thomas Hain, J Kadlec, Vasilis Karaiskos, et al. 2005. The AMI meeting corpus. In *Proceedings of the 5th international conference on methods and techniques in behavioral research*, volume 88, page 100. Citeseer.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

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: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI meeting recorder dialog act (mrda) corpus. Technical report, INTERNATIONAL COMPUTER SCIENCE INST BERKELEY CA.

J. Ulrich, G. Murray, and G. Carenini. 2008. A publicly available annotated corpus for supervised email summarization. In *AAAI08 EMAIL Workshop*, Chicago, USA. AAAI.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. MediaSum: A Large-scale Media Interview Dataset for Dialogue Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5927–5934, Online. Association for Computational Linguistics.

A Sample Data

Table 5 contains examples of conversations and summaries from the ForumSum dataset.

Ttechhunter: Im shooting a carbon element and I have the QAD Hoyt rest on it. I cannot adjust the windage of the rest without taking the rest off. Anyone have this problem? They make an Allen wrench that will fit between the riser and the rest? Thanks **splitbeam145:** they do have a short allen wrench that will fit it. Most shops will have several laying around. **BOWtechnicianTX:** those wrenches are everywhere at our shop. throw away at least 5 a day. Your local shop should give you one. **Csmith52779:** what both these guys said. That's where I got mine from was a local pro shop. **Sideler:** I took a grinder to one and made my own. **BigFoot:** The rest should have came with one in the box. Email QAD and I'm sure they will send you one for free
Ttechhunter is shooting a carbon element and needs help adjusting his rest. splitbeam145 says you can use the short allen wrench, and BOWtechnicianTX and Csmith52779 agree. Sideler says use a grinder and BigFoot says to email the company for a replacement.
UBERS4: So I had my engine covers painted - (:p) **AudiTechS4**: the upper cowl should be black in my opinion, and i would have gone red to match the other accents. doesn't look bad though **RAudi Driver**: Red would have been a good call. Props to you for doing a mod that I haven't seen yet. **Monchichi8**: nice. trying something new. **MOFSTEEL**: Reminds me of another car I saw on here not long ago. **zachf88**: I painted the washerfluid-tank cover and the upper cowl in daytona grey I'll try to get a pic of it uploaded but it's not very noticeable! **jerrym**: looks pretty good. **skiS4fun**: Nice Job, I like the white. **zachf88**: got one uploaded I'm planning on doing my airbox inlet this summer! **ny02s4**: now that looks hott! **ToMMyRsK04**: yeah red or carbon fiber woulda been my route looks good anyway
UBERS4 painted their engine covers and wants reactions. ToMMyRsK04, RAudi Driver, and AudiTechS4 conclude that it should have been a different color.

Table 5: Random examples from ForumSum dataset.

B MTurk Template

Here're instructions that were shown to the MTurk workers.

Write a summary of the conversation

Read the conversation and write a short summary.

- Be **concise**. Only cover main ideas and topics. Don't recite every message in the conversation. Try to fit the summary into 1-3 short sentences unless the conversation is long and there're multiple subjects discussed.
- Be specific. The summary must contain the **main outcome** of the conversation, not just the topic.
  - **Good:** "Jack can't install Windows 7 because of a broken license. Ann provided a working security key and Bob gave instructions for the update."
  - **Bad:** "Several users provide Jack with help troubleshooting his computer issues."
- Use **third-person** form e.g.
  - **Good:** "Ann likes oranges"
  - **Bad:** "I like oranges"
- Prefer **usernames** instead of common words like "user" and "people". Spell usernames as they are spelled in the conversation.
  - **Good:** "Ann asked"
  - **Bad:** "A user asked"
- Avoid words that don't add meaning
  - **Good:** "Ann and John discuss..."
  - **Bad:** "This seems to be a conversation where people discuss"
- Be objective. Avoid judgemental comments.
  - **Good:** "Ann and John make jokes about"
  - **Bad:** "Ann and John make stupid jokes about"
- The summary must be **grammatically correct**. Start sentences with a capital letter and use punctuation marks.
- The conversation might contain some **unknown terminology**. That's okay. Try figuring out what the conversation is about or **google the words you don't know**.

C Forums statistics

See Table 6 for the most frequent message board websites used to build ForumSum dataset. Full list of websites with counts is available at https://pastebin.com/w6wUDQx3.

D Experiment Hyper-parameters

See Table 7 for all hyperparameters we used in the experiments.

D.1 Transfer Study

For Ubuntu and NYC datasets we select 50 random examples into the validation sets and leave the other 50 in the training set. For BC3 we select 13 emails threads into validation set and use the other 27 emails threads as a training set.

We report validation numbers for all models.

D.2 Pretraining Study

For baseline pretraining experiments we used a mixture of C4 and HugeNews datasets. See (Zhang et al., 2020) for more details about these datasets.

E Human Evaluation

Overall scores are calculated by weighted average of all categories assigning scores to the categories: much better (3), better (2), slightly better (1), same (0), slightly worse (-1), worse (-2), much worse (-3). The higher the overall score is, the better ForumSum summaries are compared to SAMSum.

URL	Count
bmxmuseum.com	104
www.camaro5.com	98
metaldetectingforum.com	94
www.goldderby.com	85
discussions.texasbowhunter.com	72
linustechtips.com	72
www.camaro6.com	72
csnbbs.com	66
www.defender2.net	63
saintsreport.com	61
www.ft86club.com	60
www.neowin.net	55
www.growtopiagame.com	53
pregame.com	52
forums.thetechnodrome.com	52
www.ign.com	52
www.bbcboards.net	51
bbs.chinadaily.com.cn	51
forum.dd-wrt.com	51
wrongplanet.net	49
forums.1911forum.com	48
www.homebrewtalk.com	48
www.audizine.com	47
stargazerslounge.com	46
forums.operationsports.com	46
www.irv2.com	45
forum.woodenboat.com	45
forums.gunboards.com	44
www.calguns.net	44
mhhauto.com	43
tt.tennis-warehouse.com	43
www.visajourney.com	43
www.hltv.org	42
www.birdforum.net	42
www.democraticunderground.com	42
forums.gentoo.org	41
myanimelist.net	40
www.ar15.com	39
forums.tomshardware.com	39
www.rootschat.com	39
www.fasttech.com	38
arstechnica.com	38
www.l-camera-forum.com	38
www.jalopyjournal.com	37
forums.whirlpool.net.au	36
creditboards.com	36
forums.windowscentral.com	35
www.hmfckickback.co.uk	35
www.greatwarforum.org	35
community.betfair.com	35

Table 6: ForumSum 50 most frequent forums.

Parameter	Value
Total parameters	223M
Learning rate	1e-4
Dropout rate	0.1
Label smoothing	0.1
Batch size	1024
Max input tokens	512
Max target tokens	128
Beam size	5
Beam alpha	0.8

Table 7: Hyperparameters used in all experiments.

Dataset	Examples
ForumSum	197
SAMSum	815
NYC	94
Ubuntu	97

Table 8: Number of side-by-side pairs for each evaluation dataset. Each pair was rated by three different workers.