



Search models, datasets, users...



Datasets: knkarthick/ **samsum** like 3

Tasks: Summarization Languages: English Multilinguality: monolingual
Size Categories: 10K<n<100K Language Creators: expert-generated
Annotations Creators: expert-generated Source Datasets: original
ArXiv: arxiv:1911.12237 Tags: conversations-summarization Croissant
License: cc-by-nc-nd-4.0

Dataset card Viewer Files Community 2

Downloads last month 114

Use in Datasets library

Edit dataset card

Papers with Code



Size of downloaded dataset files:
10.3 MB

Size of the auto-converted Parquet files:
6.68 MB

Number of rows:
16,369



Dataset Viewer

Auto-converted to Parquet




API

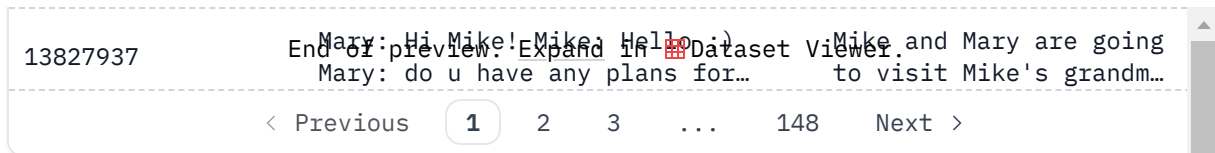
View in Dataset Viewer

Split (3)
train · 14.7k rows



Search this dataset

id string · lengths  8 10	dialogue string · lengths  29 5.47k ∅	summary string · lengths  3 300
13818513	Amanda: I baked cookies. Do you want some? Jerry: Sure! Amanda:...	Amanda baked cookies and will bring Jerry...
13728867	Olivia: Who are you voting for in this election? Oliver: Liberals...	Olivia and Olivier are voting for liberals in...
13681000	Tim: Hi, what's up? Kim: Bad mood tbh, I was going to do lots of...	Kim may try the pomodoro technique...
13730747	Edward: Rachel, I think I'm in ove with Bella.. rachel: Dont sa...	Edward thinks he is in love with Bella. Rache...
13728094	Sam: hey overheard rick say something Sam: i don't know what...	Sam is confused, because he overheard...
13716343	Neville: Hi there, does anyone remember what date I got married...	Wyatt reminds Neville his wedding anniversar...
13611672	John: Ave. Was there any homework for tomorrow? Cassandra: hello :...	John didn't show up for class due to some work...
13730463	Sarah: I found a song on youtube and I think you'll like it James...	Sarah sends James an instrumental song he...
13809976	Noah: When and where are we meeting? :) Madison: I thought...	Noah wants to meet, he quit his job, because...
13809912	Matt: Do you want to go for date? Agnes: Wow! You caught me out...	Matt invites Agnes for a date to get to know...
13727633	Lucas: Hey! How was your day? Demi: Hey there! Demi: It was...	Demi got promoted. She will celebrate that...
13729168	Mark: I just shipped the goods Mark: Tomorrow I'll send you the...	Mark just shipped the goods and he will send...
13864825	Anita: I'm at the station in Bologna Jenny: No problems so...	Anita is at Bologna station.
13729567	Leon: did you find the job yet? Arthur: no bro, still unemployed...	Arthur is still unemployed. Leon sends...
13864634	Macca: i'm so exited today Adrien: why? Macca: I've never...	Macca has done ice climbing for the first...
13815560	Isabella: fuck my life, I'm so not able to get up to work today...	Isabella feels bad after the Christmas...
13731403	Tina: I'd only like to remind you that you owe me 50 bucks Lucy: O...	Lucy owes Tina 50 dollars. She made a...
13729191	Betty: Please remind me next time that too much wine isn't good fo...	Betty feels remorse she got drunk last night...



[Dataset Card for SAMSum Corpus](#)

[Dataset Description](#)

[Links](#)

- **Homepage:** <https://arxiv.org/abs/1911.12237v2>
- **Repository:** <https://arxiv.org/abs/1911.12237v2>
- **Paper:** <https://arxiv.org/abs/1911.12237v2>
- **Point of Contact:** <https://huggingface.co/knkarthick>

[Dataset Summary](#)

The SAMSum dataset contains about 16k messenger-like conversations with summaries. Conversations were created and written down by linguists fluent in English. Linguists were asked to create conversations similar to those they write on a daily basis, reflecting the proportion of topics of their real-life messenger conversations. The style and register are diversified - conversations could be informal, semi-formal or formal, they may contain slang words, emoticons and typos. Then, the conversations were annotated with summaries. It was assumed that summaries should be a concise brief of what people talked about in the conversation in third person. The SAMSum dataset was prepared by Samsung R&D Institute Poland and is distributed for research purposes (non-commercial licence: CC BY-NC-ND 4.0).

[Languages](#)

English

Dataset Structure

Data Instances

SAMSum 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. The first instance in the training set: `{'id': '13818513', 'summary': 'Amanda baked cookies and will bring Jerry some tomorrow.', 'dialogue': 'Amanda: I baked cookies. Do you want some?\r\nJerry: Sure!\r\nAmanda: I'll bring you tomorrow :-)'}`

Data Fields

- `dialogue`: text of dialogue.
- `summary`: human written summary of the dialogue.
- `id`: unique file id of an example.

Data Splits

- `train`: 14732
- `val`: 818
- `test`: 819

Dataset Creation

Curation Rationale

In paper: 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.

🔗 Who are the source language producers?

linguists

🔗 Who are the annotators?

language experts

🔗 Annotation process

In paper: 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.

🔗 Licensing Information

non-commercial licence: CC BY-NC-ND 4.0

🔗 Citation Information

```
@inproceedings{gliwa-etal-2019-samsum,
  title = "{SAMS}um Corpus: A Human-annotated Dialogue Dataset for",
  author = "Gliwa, Bogdan and
    Mochoł, Iwona and
    Biesek, Maciej and
    Wawer, Aleksander",
  booktitle = "Proceedings of the 2nd Workshop on New Frontiers in
```

```
month = nov,  
year = "2019",  
address = "Hong Kong, China",  
publisher = "Association for Computational Linguistics",  
url = "https://www.aclweb.org/anthology/D19-5409",  
doi = "10.18653/v1/D19-5409",  
pages = "70--79"  
}
```

[🔗](#) Contributions



Company

[TOS](#)[Privacy](#)[About](#)[Jobs](#)

Website

[Models](#)[Datasets](#)[Spaces](#)[Pricing](#)[Docs](#)

© Hugging Face