WikiDialog-OQ

Dataset: github.com/google-research/dialog-inpainting

WikiDialog is a large dataset of *synthetically generated* information-seeking conversations. Each conversation in the dataset contains two speakers grounded in a passage from English Wikipedia: one speaker's utterances consist of exact sentences from the passage; the other speaker is generated by a large language model. The dataset provides the largest extant *conversational* question answering dataset. In the associated paper, we show that the quality of the questions in this dataset are comparable to manually collected datasets, and that pretraining on this dataset leads to significant quality improvements in conversational retrieval systems.

Data Card

DATASET TEAM(S)

Dialog Inpainting

DATASET CONTACT

- Group Email: dialog-inpainting-core@google.com
- Website: www.google.com

DATASET AUTHORS

- Zhuyun Dai
- Arun Tejasvi Chaganty
- Vincent Zhao
- Aida Amini
- Mike Green
- Qazi Mamunur Rashid
- Kelvin Guu

PRIMARY DATA MODALITY

Image Data
Text Data
Tabular Data
Audio Data
Video Data
Time Series
Graph Data
Geospatial Data
Multimodal (Please specify)
Others (please specify)
Unknown

DATASET SNAPSHOT

Size of dataset	30GB
Number of Instances	11,377,951
Number of Fields	6
Labelled Classes	N/A
Number of Labels	N/A
Average labels per instance	N/A
Algorithmic Labels	N/A
Human Labels	N/A
Other	N/A
Number of dialog turns	56,093,226

DESCRIPTION OF CONTENT

Each datapoint is a conversation between two speakers grounded in a passage from English Wikipedia: one speaker's utterances consist of exact sentences from the passage; the other speaker is generated by a large language model. The dataset contains the WikiDialog-OQ split described in the paper.

DATASET SUBJECT

EXAMPLE: DATA POINT

DATA FIELDS

Sensitive Data about people

Non-Sensitive Data about people

Data about natural phenomena

Data about places and objects

Synthetically generated data

Data about systems or products and their behaviors

Unknown

Others*

(*please specify)

This is what a typical data point looks like. For completeness, the very first utterance in `utterances` includes the "prompt" or role-setting utterance provided to the dialog inpainter; it can be omitted in down-stream applications.

E.g. of Data Point:

```
{'pid': '894686@1',
 'title': 'Mother Mary Alphonsa',
 'passage': 'Two years after Nathaniel's death in
1864, Rose was enrolled at a boarding school run
by Diocletian Lewis in nearby Lexington,
Massachusetts; she disliked the experience. After
Nathaniel's death, the family moved to Germany
and then to England. Sophia and Una died there in
1871 and 1877, respectively. Rose married author
George Parsons Lathrop in 1871. Prior to the
marriage, Lathrop had shown romantic interest in
Rose's sister Una. Their brother...',
 'sentences': ['Two years after Nathaniel's death
in 1864, Rose was enrolled at a boarding school
run by Diocletian Lewis in nearby Lexington,
Massachusetts; she disliked the experience.',
'After Nathaniel's death, the family moved to
Germany and then to England.', 'Sophia and Una
died there in 1871 and 1877, respectively.
Rose married author George Parsons Lathrop in
1871.', 'Prior to the marriage, Lathrop had shown
romantic interest in Rose's sister Una.',
 ...],
 'utterances': ["Hi, I'm your automated
assistant. I can answer your questions about
Mother Mary Alphonsa.",
"What was Mother Mary Alphonsa's first
education?",
 "Two years after Nathaniel's death in 1864, Rose
was enrolled at a boarding school run by
Diocletian Lewis in nearby Lexington,
Massachusetts; she disliked the experience.",
"Did she stay in the USA?",
"After Nathaniel's death, the family moved to
Germany and then to England.",
"Why did they move?",
"Sophia and Una died there in 1871 and 1877,
respectively.", ...],
 'author_num': [0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0,
1, 0],
}
```

- pid (string): a unique identifier of the passage that corresponds to the passage ids in the public OR-QuAC dataset.
- title (string): Title of the source Wikipedia page for `passage`
- passage (string): A passage from English Wikipedia
- sentences (list of strings): A list of all the sentences that were segmented from `passage`.
- utterances (list of strings): A synthetic dialog generated from `passage` by our Dialog Inpainter model. The list contains alternating utterances from each speaker ([utterance_1, utterance_2, ..., utterance_n]). In this dataset, the first utterance is a "prompt" that was provided to the model, and every alternating utterance is a sentence from the passage.
- author_num: a list of integers indicating the author number in "text". [utterance_1_author, utterance_2_author, ..., utterance_n_author]. Author numbers are either 0 or 1.

Note that the dialog in `text` only uses the first 6 sentences of the passage; the remaining sentences are provided in the `sentences` field and can be used to extend the dialog.

DATASET PURPOSE(S)	KEY DOMAINS OR APPLICATION(S)	PRIMARY MOTIVATION(S)

Monitoring **Research** Production Others (please specify)	Domains Natural Language Processing, Question Answering, Conversational AI Problem Space Conversational question answering	• Provide a large dataset of information-seeking conversations for nearly every topic in English Wikipedia. • Improve the state of the art in conversational question answering systems.
DATASET USAGE	**INTENDED AND/OR SUITABLE USE CASE(S)**	**UNSUITABLE USE CASE(S)**
Safe for production use **Safe for research use** Conditional use- some unsafe applications Only approved use Others (please specify)	• Training conversational question answering and retrieval systems.	• The dataset was created in accordance with Google's AI Principles and is not intended to be used in a way that would cause or likely to cause overall harm.
SAFETY OF USE WITH OTHER DATA	**ACCEPTABLE TRANSFORMATIONS**	**BEST PRACTICES FOR JOINING OR AGGREGATING WITH DATASET**
Safe to use with other data Conditionally safe to use with other data Should not be used with other data Unknown Others* (Please specify)	**Joining with other datasets** **Subsampling and splitting** **Filtering** **Joining input sources** Cleaning missing values **Anomaly detection** **Grouping and summarizing** **Scaling and reducing** **Statistical transformations** Redaction or Anonymization Others (please specify)	The dataset includes the title of the Wikipedia article the passage was taken from to aid users in joining with other Wikipedia-derived sources. Additionally, it uses the same unique identifier for passages as was used in the public OR-QuAC dataset to make it easier to compare and reproduce results with that dataset.
VERSION STATUS	**DATASET VERSION**	**MAINTENANCE PLAN**
Regularly Updated New versions of the dataset have been or will continue to be made available. Actively Maintained No new versions will be made available, but this dataset will be actively maintained, including but not limited to updates to the data. **Limited Maintenance** The data will not be updated, but any technical issues will be addressed. Deprecated This dataset is obsolete or is no longer being maintained.	**Current Version** 1.0.0 **Last Updated** 02/2022 **Release Date** 04/2022	• **Versioning:** N/A - WikiDialog is a static dataset. Minor releases correspond to any errors fixed in the dataset. • **Update:** WikiDialog is not updated. • **Errors:** Please contact dialog-inpainting-core@. • **Feedback:** Please contact dialog-inpainting-core@.

ACCESS POLICY	RETENTION POLICY	WIPEOUT POLICY
WikiDialog is an open-access public dataset.	N/A (Public data exemption)	N/A (Public data exemption)

DATA COLLECTION METHODS	DATA SOURCES	DATA COLLECTION
API **Artificially Generated** Crowdsourced - Paid Crowdsourced - Volunteer Vendor Collection Efforts Scraped or Crawled Survey, forms or polls Taken from other existing datasets Unknown To be determined Others (please specify)	OR-QuAC Retrieval Corpus **Passages:** Passages, titles and their identifiers in the dataset are taken from the public OR-QuAC dataset without any further processing. The passages in the OR-QuAC dataset themselves come from an English Wikipedia dump data 2019/10/20. **Date of Collection:** Oct 2019 **Instrumentation:** Scrapped **Data Modality:** Text Data Dialog Inpainting **Dialog inpainting:** Dialog inpainting generates missing utterances in a dialog using a large language model initialized from a public T5-XXL checkpoint and fine-tuned on a combination of public dialog data (see this paper for details) and two existing conversational QA datasets, OR-QuAC and QReCC. **Date of Collection:** Oct 2021 – Dec 2021 **Instrumentation:** Large language model **Data Modality:** Text Data	Dialog Inpainting **Collected and included** • sentences: Sentences extracted from the passage using the Google Cloud NL API. • text: Utterances generated by the dialog inpainting model. • author_num: Alternating author numbers for the utterances in `text`.

INCLUSION CRITERIA	EXCLUSION CRITERIA	DATA PROCESSING
Dialog Inpainting We used all passages in the OR-QuAC retrieval corpus and included all the data generated by the model in the dataset.	Dialog Inpainting N/A (no data filtering was applied)	Dialog Inpainting N/A (no data processing was applied)

SENSITIVE DATA	FIELDS WITH SENSITIVE DATA	SECURITY AND PRIVACY HANDLING
User Content User Metadata User Activity Data Identifiable Data S/PII Business Data Employee Data Pseudonymous Data Anonymous Data Health Data Children's Data **None** Others* (*please specify)	N/A	N/A

SENSITIVE HUMAN ATTRIBUTES	SOURCE(S) OF HUMAN ATTRIBUTES	RATIONALE FOR COLLECTING HUMAN ATTRIBUTES
Race Gender Ethnicity Socio-economic status Geography Language Sexual Orientation Religion Age Culture Disability Experience or Seniority **None** Others (please specify)	N/A	N/A

ML APPLICATION(S)	EVALUATION - RESULTS	EVALUATION - PROCESS

Question answering, Information retrieval, Conversational modeling	We use this dataset to *pretrain* conversational passage retrieval systems and fine-tune them on a target dataset. Please see the paper for details on the evaluation methods and extended results. T5-Base dual encoder, pretrained on WikiDialog-OQ, fine-tuned and evaluated on OR-QuAC Mean reciprocal rank@5 65.3 T5-Large dual encoder, pretrained on WikiDialog-OQ, fine-tuned and evaluated on QReCC Mean reciprocal rank 58.9	Please refer to the paper for extended details on the evaluation process.
	MODEL DESCRIPTION(S) AND STATS	**EXPECTED PERFORMANCE & KNOWN CAVEATS**
	T5-Large dual encoder Initialize a dual encoder model using a public T5-Large checkpoint, pre-train the model on WikiDialog and fine-tuning on a target conversational question answering dataset. In each setting, we train the model to score the target passage more highly than a random set of "negative" passages. Please see the paper for details on how the negative passages were sampled.	Conversational Retrieval (without Reranking) **Expected performance:** We find significant improvements on finetuned retrieval performance when using WikiDialog to pre-train a conversational retriever. See paper for more details and complete results. The results below use mined hard negatives, as is standard practice. • **QReCC** • T5-Base Baseline (no pretraining): 53.4 MRR • T5-Base + WikiDialog-OQ Pretraining: 58.9 MRR • **OR-QuAC** • T5-Base Baseline (no pretraining): 53.6 MRR • T5-Base + WikiDialog-OQ Pretraining: 65.3 MRR **Known Caveats:** • Using mined hard negatives in the fine-tuning stage significantly improves retrieval performance. In contrast, we find that using mined hard negatives in the pre-training stage does not improve and sometimes hurts overall performance.

Terms of Art

Concepts and Definitions referenced in this Data Card

Text Retrieval Models	Mean Reciprocal Rank (MRR)	T5
Definition: A retrieval model takes a user query q and retrieves N documents from a large document corpus that are most likely to answer the user query. Source: Wikipedia	Definition: a metric to measure the quality of a retrieval model. For Q queries, let $rank_i$ be the rank of the correct answer in the returned results, then:	Definition: a text-to-text transformer model. Source: Paper, Google Blog

Interpretation: We evaluate WikiDialog quantitatively by using it to improve *conversational* retrieval models. These are retrieval models that additional use the conversation history—previous queries and results—when retrieving documents.	$$\text{MRR} = \frac{1}{\lvert Q \rvert} \sum_{i=1}^{\lvert Q \rvert} \frac{1}{\text{rank}_i}.$$ Source: Wikipedia	

Reflections on Data

Potentially sensitive language in the dataset	In the paper, we provide a quantitative estimate for how often the dataset could contain potentially sensitive language that may perpetuate unfair biases. We approached the problem by curating a set of 700 terms from the literature (Bolukbasi et al., 2016; Garg et al., 2018; May et al., 2019; Nadeem et al., 2020; Abid et al., 2021) related to sensitive characteristics—such as race, ethnicity, gender, and sexual orientation. Many instances of these terms are well-motivated: for example, a dialog from a passage about transgender rights in Canada includes the question "What does anti-discrimination act mean in relation to transgender people?" We further refined the approach to instead look at co-occurrences between these terms and adjectives that may have negative connotations, focusing on instances where the terms were not explicitly mentioned in the passage. We find that 0.2–0.5% of dialogs in the dataset contain such potentially sensitive interactions, but it is difficult to establish if they perpetuate unfair bias without expert manual review. Therefore, we advise users to note these observations and exercise care while using the dataset. Further details of our approach and examples can be found in the paper.