Datasheet for ArgSciChat

1 Motivation for Datasheet Creation

Why was the dataset created?

We collected ArgSciChat to overcome the lack of dialogue systems in the scientific domain, where dialogues are characterized by exchanges of information and opinions grounded on a scientific paper. Furthermore, dialogues in ArgSciChat follow a set of intents designed to convey these two types of interactions between dialogue partners. Thus, ArgSciChat involves different dynamics of conversational agents, such as goal-oriented dialogue, document grounding, and interactions with multiple intents.

What (other) tasks could the dataset be used for?

In addition to the tasks of *response generation* and *rationale selection* discussed in this work, ArgSciChat contains sentence-level annotations regarding our intent categories: Exploratory (EXP) and Argumentative (ARG). These annotations could be used for *intent classification*, i.e., the task of determining the intent of a sentence and developing advanced dialogue agents for response generation. For instance, intent labels could be given as input to an agent that generates a message by generating a sentence at a time. Furthermore, ArgSciChat could be used to define a dialogue agent that plays as the proponent `P` to emulate an interested researcher. Lastly, the ARG category of intents could be considered in the context of Opinion Mining or Argument Mining.

Has the dataset been used already?

We introduce ArgSciChat in this work. Thus, no other work has used ArgSciChat for research at the time of writing.

Who funded the creation of the dataset

We will provide this information in case of paper acceptance to not violate the authors' anonymity.

2 Dataset Composition

What are the instances?

Dialogues in ArgSciChat are turn-based. `P` and `E` generate a dialogue by alternating messages. We denote as dialogue turn a pair of subsequent messages in a dialogue.

For the task of response generation, we evaluate a dialogue agent that takes the `E` role. In this scenario, a dialogue agent receives a `P` message as input, denoted as a query Q, and generates the corresponding `E` message as a reply. Thus, each dialogue turn in ArgSciChat is an instance for response generation. We also consider other inputs in addition to a query Q: (a) the scientific paper (P) `E` and `P` discuss; (b) all messages exchanged before Q, also known as dialogue history (H); (c) the reference rationales (R) of the `E` message. In our work, we evaluate three input configurations for a dialogue agent:

1. (Q, P): the agent receives a query Q and the scientific paper P as input.

2. (Q, P, H): the agent receives the dialogue history H in addition to a query Q and the scientific paper P.

3. (Q, R): the agent receives a query Q and the reference rationales of the `E` message that replies to Q.

Regarding supportive rationale selection, the instances and the input configurations are the same as response generation. In particular, we introduce supportive rationale selection as an auxiliary task for response generation. In our work, a dialogue agent is trained to address these tasks jointly. Unlike response generation, a dialogue agent has to predict which sentences in P are the reference rationales of a `E` message.

How many instances are there?

ArgSciChat contains 41 dialogues about 20 scientific papers in the Natural Language Processing (NLP) domain. In total, ArgSciChat contains 249 dialogue turns (498 messages and 1034 sentences). Dialogues were created by 23 NLP researchers that agreed to participate in our study.

Is everything included or does the data rely on external resources?

ArgSciChat does not rely on external resources. We provide the dialogues, the paper content on which dialogues are based, and the intent annotations. The dataset is available at `https://github.com/UKPLab/acl2023-argscichat`.

Are there recommended data splits or evaluation measures?

We carry out a five-fold cross-validation routine for evaluating a dialogue agent on response generation and rationale selection tasks. Folds are defined by splitting dialogues at the scientific paper level. Thus, splits do not share instances, i.e., dialogue turns, belonging to the same dialogue. We release the folds and the random seeds used for reproducibility and comparison.

Regarding evaluation measures, we consider a standard set of metrics. For rationale selection, we consider a sentence-level F1-score (Rationale-F1). A dialogue agent has a classification layer that predicts 1 if it reputes that a sentence is a reference rationale for a `E` message, 0 otherwise. We consider a token-level F1-score between a generated response and its corresponding `E` message (Message-F1) for response generation. This metric was used in other well-known datasets like SQUAD and QASPER. Additionally, we consider transformer-based metrics like BertScore and MoverScore.

3 Data Collection Process

How was the data collected?

Dialogues in ArgSciChat were collected using our proposed methodology and corresponding implementation. Subjects had freedom of choice regarding scientific paper selection. We did not impose any limitations on paper or topic selection. Our implementation supports an automatic paper retrieval functionality based on a subject's Google Scholar profile. In our implementation, only publicly available scientific papers that did not involve a paywall were considered. Alternatively, subjects had the opportunity to submit the PDF hyperlink of their preferred papers manually. Our methodology consists of four major steps. We ask for subjects' consent, to provide contact information, including their research profile, and to select a few papers on which they are experts. Then, subjects propose time slots at which they take the role of `E` in a dialogue. Subjects are presented with a timetable reporting the time slots selected by all subjects. We ask subjects to select a few time slots according to their preference. In these slots, subjects take the role of `P` in a dialogue. Before conducting any dialogue, subjects learn about our dialogue formulation. Subjects join their scheduled time slots. Once both subjects are present, the dialogue begins.

Who was involved in the data collection process?

We (the authors) carried out initial pilot studies with the proposed methodology to refine the data collection process and guidelines. Subsequently, we configured our implementation for collecting dialogues from the pool of subjects. We invited senior and junior researchers from two large NLP groups in Europe as a candidate pool for participation. Each subject played as `E` and `P` depending on their scheduled time slots as described in our work.

Over what time-frame was the data collected?

The data was collected over two weeks.

Does the dataset contain all possible instances?

No. Dialogues in ArgSciChat regard a few NLP topics. Dialogues on scientific papers can be conducted over any topic or research domain. We chose the NLP domain for dialogues since we (the authors) have expertise in this domain. This choice also facilitated the definition of participation pool through our research network. Furthermore, we limited the paper's content to abstract and introduction sections to reduce `E`'s effort while providing enough information to sustain a dialogue. Thus, we do not observe interactions between subjects in other sections of a paper and related discussions.

If the dataset is a sample, what is the population?

The dialogues in ArgSciChat concern a limited number of scientific papers in the NLP domain. Furthermore, dialogues are limited to the abstract and introduction sections of a paper. For these reasons, dialogues in ArgSciChat are only a small sample of the possible dialogues grounded on a scientific paper. Different factors and simplifications have to be taken into account: (a) the topic of a paper; (b) the common background of subjects; (c) the available content of a paper; and (d) the dialogue setting (e.g., in our setting, dialogues had a time limit which restricted the number of interactions between subjects).

4 Data Preprocessing

What preprocessing/cleaning was done?

Scientific papers were automatically converted from PDF format to textual format via GROBID (https://github.com/kermitt2/grobid). Dialogues and the scientific papers were tokenized using NLTK (https://www.nltk.org/). Additionally, we filtered out dialogues that ended abruptly or had less than six dialogue turns.

Was the "raw" data saved in addition to the preprocessed/cleaned data?

We include the full text of dialogues and associated scientific papers. Additionally, we provide a version of the dataset that is ready to use for reproducing our experimental results.

5 Dataset Distribution

How is the dataset distributed?

The dataset is available at https://github.com/UKPLab/acl2023-argscichat.

When will the dataset be released/first distributed?

The dataset is already ready to use. We will release the dataset in case of paper acceptance.

What license (if any) is it distributed under?

ArgSciChat is distributed under the MIT license.

6 Legal & Ethical Considerations

Were participants told what the dataset would be used for and did they consent?

We ask experts to read and confirm a consent concerning data privacy and informed consent before signing up for our tool. In the form, we explicitly state the aim of the study and the later use of collected data. We provide detailed information to the subjects about the personal data information we require for participation and its temporary usage throughout the study. Subjects can request data deletion at any given step of the study. All subjects who agree to sign-up also consent to participate in the study.

If it relates to people, could this dataset expose people to harm or legal action?

The collected dialogues are anonymized and do not contain personal information about the subjects or information that could reveal their identity. Furthermore, the scientific papers that were used for dialogues were publicly available. Our data collection process and how subjects were selected and instructed to collect dialogues avoided situations where harmful or irrelevant messages could be written in a dialogue. The dataset could contain biases concerning `E` messages that have ARG intent. In particular, subjects that played as `E` were experts about the content of the scientific papers discussed in a dialogue. This might lead to argumentative interactions, e.g., rebuttals, that strongly defend the statements remarked in the discussed scientific paper without considering other points of view.