A Dataset of Argumentative Dialogues on Scientific Papers

Federico Ruggeri[1 *] **Mohsen Mesgar**[2 †] **Iryna Gurevych**[3]

[1]DISI, University of Bologna
[2]Bosch Center for Artificial Intelligence, Renningen, Germany
[3]Ubiquitous Knowledge Processing Lab (UKP Lab),
Department of Computer Science and Hessian Center for AI (hessian.AI),
Technical University of Darmstadt

[1]federico.ruggeri6@unibo.it [2]mohsen.mesgar@bosch.com
[3] www.ukp.tu-darmstadt.de

Abstract

With recent advances in question-answering models, various datasets have been collected to improve and study the effectiveness of these models on scientific texts. Questions and answers in these datasets explore a scientific paper by seeking *factual* information from the paper's content. However, these datasets do not tackle the *argumentative* content of scientific papers, which is of huge importance in persuasiveness of a scientific discussion. We introduce ArgSciChat, a dataset of 41 argumentative dialogues between scientists on 20 NLP papers. The unique property of our dataset is that it includes both exploratory and argumentative questions and answers in a dialogue discourse on a scientific paper. Moreover, the size of ArgSciChat demonstrates the difficulties in collecting dialogues for specialized domains. Thus, our dataset is a challenging resource to evaluate dialogue agents in low-resource domains, in which collecting training data is costly. We annotate all sentences of dialogues in ArgSciChat and analyze them extensively. The results confirm that dialogues in ArgSciChat include exploratory and argumentative interactions. Furthermore, we use our dataset to fine-tune and evaluate a pre-trained document-grounded dialogue agent. The agent achieves a low performance on our dataset, motivating a need for dialogue agents with a capability to reason and argue about their answers. We publicly release ArgSciChat[1].

1 Introduction

The enormous and ever-growing number of scientific papers (Munroe, 2013; Ronzano and Saggion, 2015) make scientific text processing imperative. A considerable body of research in NLP is dedicated to developing methods to provide insights from such texts (Beltagy et al., 2019; Zhang et al., 2021; Wadden et al., 2022; Zhang et al., 2021; Parveen et al., 2016; Mysore et al., 2022). Recent advances in Question-Answering (QA) models made them a great asset in accessing the content of scientific papers (Dasigi et al., 2021). To develop and evaluate such QA models, various datasets (Section 2) have been collected (Dasigi et al., 2021; Pampari et al., 2018; Pappas et al., 2018; Jin et al., 2019; Tsatsaronis et al., 2015; Krallinger et al., 2020). For each scientific paper in these datasets, a set of questions with the exploratory intention is defined. These questions explore the content of the paper by seeking *factual* information. Thus, answers in these datasets are mainly limited to yes/no, named entities, or extracted from the text of the paper.

Scientific answers should be explained with argumentation to be persuasive (Gilbert, 1977). If an answer is not persuasive, it might need an argumentation to be well understood. To build models with the ability to argue about a scientific fact motivates a vital need for a dataset of scientific argumentative dialogues. A scientific argumentative dialogue includes interactions with both exploratory, EXP, and argumentative, ARG, intents. EXP intents address factual information and ARG intents aim at the argumentative content of a paper. Collecting such a dataset is challenging for two reasons. First, while there are linguistic theories about argumentative dialogues (Walton, 2008; Walton and Macagno, 2007); there is no intent set for collecting argumentative questions and answers in the scientific domain. Second, discussions in a scientist domain (and an expert domain in a broader perspective) need specialized domain expertise. So, the process of incentivizing scientists and aligning them to chat about a paper should be optimized for scientists. Thus, off-the-shelf crowdsourcing platforms, which are applied for data collection in generic domains using crowd subjects, hardly are applicable for our goals in this paper.

[*] Work done while conducting an internship at UKP Lab.
[†] Work done while being a postdoc at UKP Lab.
[1]https://github.com/UKPLab/acl2023-argscichat

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7684–7699

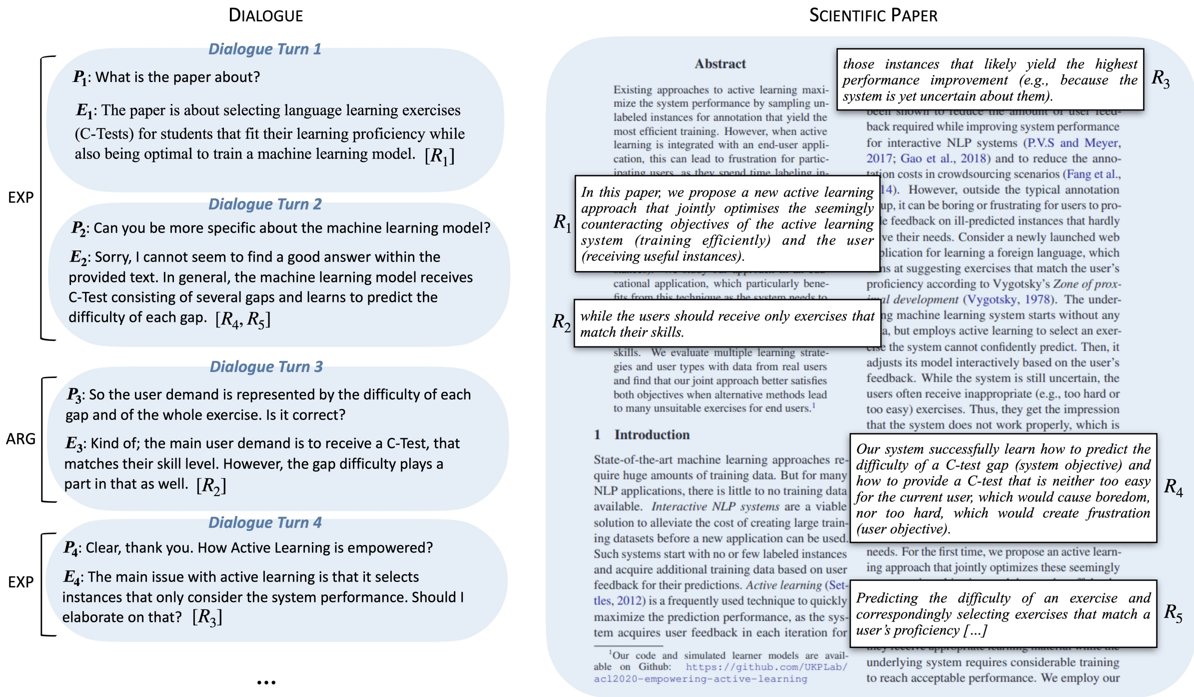


Figure 1: An example dialogue from our ArgSciChat dataset. Rationales $\{R_1, \ldots, R_6\}$ are highlighted in the scientific paper. Dialogue partners alternate between exploratory (EXP) and argumentative (ARG) intents.

In this work, we introduce *ArgSciChat*, the first dataset of text-based argumentative dialogues on NLP papers (Section 3). Figure 1 shows an example dialogue from our dataset. Dialogues are in English and conducted synchronously by *subjects* who are NLP scientists. Each dialogue is about a single paper. Each subject plays a different role in the dialogue. `E` is the **expert** on the research presented in the paper, and `P` (for **proponent**) is the subject aiming to learn about the scientific paper by posing questions to the expert. Importantly, `E` and `P` argue about the factual content of the paper and each other's opinion based on **rationale**, *R*, selected from the paper. To collect ArgSciChat, we define a set of intent classes for interactions (i.e., questions and answers). These intents address exploratory and argumentative interactions. We also develop a new dialogue collection tool in which we account for scientist' preferences during the dialogue collection process by allowing them to choose a paper to chat about.

To put our dataset in context, we compare (Section 4) it with three recent document-grounded QA datasets, i.e., CoQA (Reddy et al., 2019), QuAC (Choi et al., 2018) and Doc2Dial (Feng et al., 2020). We then evaluate the diversity of dialogues in our dataset. Finally, we use our dataset to fine-tune and evaluate LED (Beltagy et al., 2020), a pre-trained dialogue agent for answering questions about a long text in generic domains. The goal of this experiment is to study the difficulty level of our dataset for current dialogue systems. Our results (Section 5) confirm the difficulty of our dataset. Our main contributions are **(1)** a challenging scientific argumentative dialogue dataset on NLP papers, **(2)** a new set of intent classes for interactions in argumentative dialogues, **(3)** a new tool for collecting dialogues between scientists. We hope that our contributions pave the way for building human-like dialogue agents in scientific domains. These agents should foster our access to the factual content of scientific papers by answering our questions and also scientifically arguing about the paper's content.

2 Related Work

While the problem of distinguishing between opinionated and factual content is not new (Yu and Hatzivassiloglou, 2003; Kang et al., 2018; Hua et al., 2019; Cheng et al., 2020), there is no dataset consisting of dialogues with both exploratory (EXP) and argumentative (ARG) interactions in a scientific domain. The most similar datasets to ours deal with QA tasks on scientific papers (Dasigi et al., 2021; Pampari et al., 2018; Pappas et al., 2018, 2020; Jin et al., 2019; Welbl et al., 2018; Tsatsaronis et al., 2015; Krallinger et al., 2020).

Intent	Definition	Example
Exploratory (EXP)		
Ask Info (AI)	A question to seek factual information from a paper.	*What data set is used to train a classifier?*
Give Info (GI)	An answer to give factual information.	*We consider the Partial Latin Square completion problem.*
Switch Topic (ST)	A request or proposal to shift the topic.	*Which intent would you like to know more about?*
Argumentative (ARG)		
Ask Opinion (AO)	A question to ask about opinion.	*Which do you think is the strongest point of your approach?*
Give Opinion (GO)	An answer to provide an opinion.	*It is interesting that such simple techniques help.*

Table 1: We define five intent classes for questions and answers in a scientific argumentative dialogue. EXP intents are about factual information and ARG intents are about argumentative content of a paper.

These datasets include interactions with EXP intents and lack ARG interactions. Moreover, questions in these datasets weakly link to each other to resemble a dialogue. In contrast, ArgSciChat consists of coherent dialogues with both EXP and ARG interactions. Answers in the aforementioned QA datasets are limited to named entities extracted from a paper's content. Conversely, dialogues in ArgSciChat are exclusively in a free-form text, a vital property of human-like dialogues (Reddy et al., 2019).

While there are many dialogue datasets (see the survey provided by Serban et al. (2018)) in generic domains, there are a few datasets that contain dialogues grounded on (non-scientific) documents, e.g., Wikipedia. These dialogues are conducted by non-expert crowd subjects. In particular, we refer to CoQA (Reddy et al., 2019), QuAC (Choi et al., 2018), and Doc2Dial (Feng et al., 2020). Our work differs from these datasets in two ways. First, dialogues in our dataset are grounded on scientific papers , which follow a specific argumentative structure (Gilbert, 1976, 1977) and include exploratory and argumentative interactions. The other datasets, only include exploratory interactions on generic texts. Second, unlike the other datasets, dialogues in our dataset happen between scientists. It is challenging to collect synchronous dialogues between scientists given their limited time budget.

The theory of argumentative dialogues was addressed many years ago (Walton and Macagno, 2007). The arguably most prominent argumentative dialogue system is the IBM project debater (Slonim et al., 2021), which debates different topics. The utilized retrieval engine and dataset of this system are not available to the public which hinders its application in scientific domains. There are also other argumentative datasets under dialogue settings (Lawrence and Reed, 2019; Fazzinga et al., 2021) for various domains such as political debates (Mestre et al., 2021) and online rebuttals (Orbach et al., 2019). Unlike these datasets, ArgSciChat includes scientific argumentative dialogues where the arguments are grounded in the content of scientific papers.

3 Methodology

Our goal is to collect synchronous argumentative dialogues about the content of scientific papers. These dialogues should include EXP and ARG interactions (i.e., questions and answers) between scientists. Thus, we begin with defining a set of intents for interactions in scientific argumentative dialogues. We then introduce our dialogue collection methodology and a tool to collect such dialogues.

3.1 Defined Intent Set

There are no intent sets for collecting scientific argumentative dialogues. Thus, we define a set of intents for interactions in such dialogues by taking inspiration from a fundamental theory proposed by Walton (2008). It is worth noting that Walton (2008) defines various types of interactions in argumentative dialogues. By taking into account their goals, only two types of interactions are relevant to scientific dialogues: information-seeking to extract information and persuasion to discuss the claims of a scientific paper. For example, We do not include interaction types such as "personal attacks" and "appeals to emotion" because scientific discussion is supposed to be objective.

Formally, we define five intent classes for scientific argumentative dialogues. Table 1 summarizes the definitions of these intents and shows an example for each intent class. We group these fine-grained intent classes into the exploratory and argumentative groups. The exploratory (EXP) intents aim to seek or provide factual information. In particular, this group includes "Ask Info (AI)", "Give Info (GI)" and "Switch Topic (ST)". The ST intent is needed to ensure the continuity in dialogues. The argumentative (ARG) interactions aim to convey persuasion about factual information. In particular, we define two intent classes in the EXP group: "Ask Opinion (AO)" and "Give Opinion (GO)". The AO intent seeks the opinion of a dialogue partner. In a scientific dialogue, it translates to asking for scientific argumentation about some factual information discussed in prior dialogue turns. In return, the GO intent indicates dialogue sentences which aim to answer questions that ask for argumentation.

3.2 Dialogue Formulation

To provide sufficient factual and argumentative knowledge for conducting scientific argumentative dialogue, we ground a dialogue in the content of a scientific paper. Thus each dialogue is performed between two human subjects on a scientific paper. In each dialogue, each subject receives either role `E` (expert) or `P` (proponent). `P` accesses only the paper's title, whereas `E` accesses the abstract and introduction sections of the scientific paper. We limit the paper content to the abstract and introduction sections because these two sections present the gist of the paper sufficiently to sustain a dialogue.

`P` and `E` discuss the content of a scientific paper from both factual and argumentative aspects according to the intent set we defined for scientific argumentative dialogues. In particular, to collect EXP interactions, we instruct `P` to use the title to start the conversation by asking for factual information about the title. `E` answers the question using the content of given sections of the paper. Since `E`'s answers should be grounded in the content of the scientific paper, we ask `E` to select up to two text spans from the paper that are used in generating the answer. Consistent with terminology used by Reddy et al. (2019), we refer to these text spans as the rationale. We emphasize that a rationale does not imply any argumentative function of the text span (e.g., a rationale can be a claim like R_5 in Figure 1). Rationales help human subjects to ground their sentences in scientific content of a paper to conduct high-quality dialogues.

Aiming to a human-like dialogue (Reddy et al., 2019), `P` and `E` to write down their message (which could be a question or an answer) in free-form textual sentences. A pair of a `P`'s and an `E`'s messages constitutes a dialogue turn. An `E`'s message and selected rationales are displayed to `P` to initiate the next dialogue turn. `P` and `E` argue about factual information in prior dialogue turns. The explored factual information should be easier to understand through argumentation. Following Reddy et al. (2019), to limit the occurrences of unanswerable questions, `E` can switch topic using the paper's content.

3.3 Dialogue Collection Tool

To collect scientific argumentative dialogues on scientific papers, we focus on the NLP domain. This domain includes many interdisciplinary scientific papers, and also lays down a common background between our human subjects. For each dialogue we ask two human subjects who are experts in NLP to synchronously chat about factual and argumentative content of an NLP paper.

A key challenge in implementing our data collection methodology is to optimize this process for scientists. The conventional crowdsourcing frameworks (e.g., AMT and Upwork) lack three main features regarding dialogue collection in the scientific domain: (1) flexibility in scientific paper selection, (2) time slots scheduling, and (3) synchronous participation. To meet these requirements and incentivize scientists' participation, we introduce an implementation for our dialogue collection methodology. We develop a new unified web-based tool, letting human subjects register, suggest papers for dialogue collection, share their availability for conducting a synchronous dialogue, and chat in a written format about an NLP paper. While subjects sign up in our tool, they confirm a consent which authorizes the usage of their dialogues and papers for any sort of research purposes. We define the consent form by following GDPR[2] and the ethics guidelines for trustworthy AI[3].

To encourage subjects to participate in a dialogue, we let them select a few scientific papers from an automatically retrieved list of their recent publications. This idea relieves subjects' burden in reading the selected paper since they are the author of those papers. Moreover, from the subjects' perspective, participating in a dialogue about their publications is an informal advertisement for their research. To align the subjects with each other time-wise, we ask subjects with role `E` to announce multiple time slots in which they are available for presenting their selected scientific paper in our tool.

[2] https://gdpr.eu/tag/gdpr/

[3] https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1

To incentivize subjects with role P to ask scientifically deep questions and not just generic remarks, we let them choose time slots based on their time schedule and also the titles of papers associated with the slots. Appendix A reports implementation details of the tool and screenshots of its interfaces.

3.4 ArgSciChat Dataset

We used our tool to collect scientific argumentative dialogues for the ArgSciChat dataset. We invited 31 senior and junior scientists in NLP from two large NLP groups in Europe to participate in our dialogue collection study as human subjects. 23 of the invited scientists (74.2%) accepted our invitation and participated in at least one dialogue using our tool.

We collected 41 dialogues on 20 NLP papers, consisting of 498 messages. Each message is an a text generated by dialogue partners during a dialogue. We present these dialogues and their corresponding papers as the ArgSciChat dataset.

4 Experiments

We collected a set of argumentative dialogues on NLP papers. We design a set of experiments to assess if the collected dialogues reflect the properties of scientific argumentative dialogues, and then evaluate the difficulty of our dataset for an advanced dialogue agent.

4.1 Comparison with Similar Datasets

To put the ArgSciChat dataset in the context of discussed literature (Section 2), we compare it with its most similar conversational datasets. These datasets are CoQA (Reddy et al., 2019), QuAC (Choi et al., 2018), and Doc2Dial (Feng et al., 2020). It is worth noting that there exist a few QA datasets on scientific papers like QASPER (Dasigi et al., 2021). However, the lack of dialogical properties of these datasets forbids making a fair comparison with our dataset. For this reason, we only consider popular conversational datasets for comparison. They include text-based questions and answers about factual information from texts in generic domains. Therefore, this experiment reveals the similarities and differences between dialogues in a specialized domain (i.e., NLP) and those in generic domains.

In particular, we focus on the linguistic properties of dialogues. We measure the dialogue length with the number of dialogue turns. As an estimate for the dataset size, we use the number of dialogues in each dataset. For the CoQA and QuAC dataset, each set of questions and answers collected for a text is considered as one dialogue. Finally, we compute the message lengths in terms of the number of multi-sentence messages (MSM) and the average number of tokens in messages.

4.2 Analyzing Dialogues of ArgSciChat

To obtain a deep understanding of the dialogues in ArgSciChat, we analyze dialogues in this dataset from three perspectives.

- First, we assess the diversity of dialogues collected for a paper. In particular, limiting paper content may lead to similar dialogues on the same paper. To ascertain such a possibility, we experiment with a form of semantic diversity to evaluate the following dialogue dynamics: the percentage of P messages, E answers, and selected rationales that are semantically different.

- Second, we evaluate the properties of selected rationales, such as how they are used in dialogues and from which portion of papers they are chosen.

- Third, we measure to what extent the dialogues are exploratory and argumentative. In particular, we carry out an annotation study to label dialogue sentences with EXP and ARG intents and evaluate the distribution of these intent labels.

4.3 Using ArgSciChat to Evaluate Dialogue Agents

To assess the difficulty of ArgSciChat for recent dialogue agents, we synthesize a dialogue agent to take the role E. Consequently, we define two evaluation tasks for this agent:

- *Rationale selection*: For this task, the agent should select up to two sentences from a given scientific paper as rationales;

- *Answer generation*: For this task, the agent generates a free-form text to generate an answer to a P's question.

We use the LED agent (Beltagy et al., 2020) to conduct this experiment. According to Dasigi et al. (2021), LED achieves the state-of-the-art performance for these two tasks on scientific papers in

QASPER, which is a scientific QA dataset. LED is based on Longformer architecture and pre-trained to answer questions about the content of a long text in a generic topic (i.e., Wikipedia). We compare the following LED configurations: **LED(Q, P)**, **LED(Q, P, H)**, and **LED(Q, R)**. We use the term Q (query) to refer to a question expressed by a proponent. P refers to the content of an NLP paper. H (history) shows a sequence of tokens, including all messages exchanged before Q. R (rationales) is a sequence of tokens including all rationales selected by human subjects to generate reference answers. For the sake of fair comparisons, we set all training settings identical to those reported by Dasigi et al. (2021). For reproducibility, we report the values of all hyper-parameters used to train LED in Appendix C.

Because of the small number of data points, we carry out a five-fold cross-validation routine for this experiment. We create fold splits such that all dialogues about the same scientific paper are in the same split. We report the number of dialogue turns as data samples used in the training, validation, and test sets of each fold in Appendix C. For evaluation metrics, similar to Dasigi et al. (2021), for rationale selection, we compute the F1 score over candidate sentences in a scientific paper against the reference rationales. We denote this metric as *Rationale-F1*. For answer generation, we use the token-level F1 score introduced in SQUAD (Rajpurkar et al., 2016). This metric is computed over individual words between the generated answer and the reference message. We denote this metric as *Message-F1*. Moreover, we leverage advanced evaluation metrics, i.e., BERTScore (*BScore*) (Zhang et al., 2020) and MoverScore (*Mover*) (Zhao et al., 2019), for text generation. These metrics that have been shown to correlate with human evaluation.

As a baseline for rationale selection, we use a retrieval-based method. This method computes the cosine-similarity score between representations of a query Q and representations of each sentence in a paper. We use TF-IDF with default parameters in *Scikit-learn* and S-BERT (Reimers and Gurevych, 2019) to obtain the vector representations. We rank the sentences concerning their similarity scores with the query and select the top-two sentences as the rationales. As another strong baseline, we ask three NLP experts, who are different from those participated in dialogue collection, to extract rationales for a given query.

Dataset	Avg. Turns	% MSM	Avg. Length	# Dial.
CoQA	15.5	0.2%	4.7	8k
QuAC	7.3	4.0%	11.4	13.5k
Doc2Dial	6.4	17.8%	16.3	4.8k
ArgSciChat	6.3	50.8%	38.2	41

Table 2: **Avg. Turns** shows the average number of turns per dialogue. **% MSM** depicts the percentage of multi-sentence message in a dataset. **Avg. Length** indicates the average number of tokens per message. **# Dial.** indicates the number of dialogues. Rationales are not considered as being part of answers.

5 Results

5.1 Comparison with Similar Datasets

Table 2 reports the characteristics of dialogues in the compared datasets. In terms of the average number of dialogue turns, ArgSciChat is comparable with QuAC and Doc2Dial. Compared with CoQA, dialogues in ArgSciChat contain fewer dialogue turns. This is because in the limited time slots, human subjects in our study should read the content of a paper, think about the scientific argumentation and write the answers in their own words. However, in CoQA questions are short in length and answers are named entities selected from a WiKi text, which is easier-to-grasp than a scientific text. The interactions in ArgSciChat are beyond text spans and consist of free-form textual sentences. In particular, ArgSciChat contains a higher percentage (50%) of multi-sentence messages than the other datasets. This observation is also confirmed by the average message length, which shows the single-sentence messages also have higher length than messages in the other datasets.

In terms of the dataset size (# Dial.), ArgSciChat contains fewer dialogues than the other datasets. The limited number of dialogues in our dataset is symptomatic of expert domains, where the data collection process is subject to tight requirements like matching scientists' schedules. Unlike our dataset, the compared datasets do not include human-like dialogues. These datasets include contextualized question-answering interactions grounded in generic documents. Since no expert knowledge is required to explore such texts, dialogues take place between crowd contributors in these datasets. Conversely, dialogues in ArgSciChat are conducted by experts in NLP as a scientific domain.

5.2 Analyzing Dialogues of ArgSciChat

Dialogue Diversity. On average ArgSciChat contains two dialogues for each NLP paper. We study to what extent the messages (which could be questions or answers) exchanged in these dialogues are semantically diverse. To do so, we group messages in dialogues grounded in one paper into three categories: **G1:** semantically similar `P` messages; **G2:** `E` messages associated with `P` messages in G1; and **G3:** `E` messages and their corresponding rationales. For each group, we compute the semantic diversity between any pair of messages, using the cosine distance between the sentence representations obtained with S-BERT[4]. If the cosine distance between the representations of the messages in a pair is below a threshold, we consider the messages to be semantically diverse. We empirically found that a threshold value of 0.5 is sufficient for the purpose of this experiment. We report qualitatively similar results obtained with different thresholds and different embedding models in Appendix B.

We compute the average percentage of semantically diverse pairs in each group over all papers with multiple dialogues. About 90% of messages in group G1, 63% in group G2, and 43% in group G3 are semantically diverse. These results suggest that our dialogue formulation yields diverse dialogues even on identical papers. This property of ArgSciChat is important when its dialogues are used to evaluate dialogue agents (Reddy et al., 2019).

Rationales Distribution. Up to two rationales are used as groundings for answers generated by `E`. Table 3 shows that about 61.1% (129 out of 211) of these answers are grounded in one rationale and 38.9% (82 out of 211) on two rationales. These statistics show that a considerable portion of questions in ArgSciChat are difficult and need more than one rationales to be answered. For the two-rationales answers, the average sentence-based distance between the rationales in the scientific paper is 5.8. This observation indicates that the rationales in the two-rationales answers are almost in a local context, e.g., a paragraph. This makes sense because nearby sentences in a scientific paper are about a similar topic. Table 3 shows that a large portion (61%) of all the rationales are selected from the introduction section. This shows that both abstract and introduction were used by human subjects, and importantly most questions are scientifically deep such that abstract is not enough to answer them.

Property	Value
1-Rationale answers	61.1 %
2-Rationale answers	38.9 %
Avg. sentence distance between rationales	5.8
Rationales from abstract	38.6 %
Rationales from introduction	61.4 %

Table 3: Studied statistics of rationales.

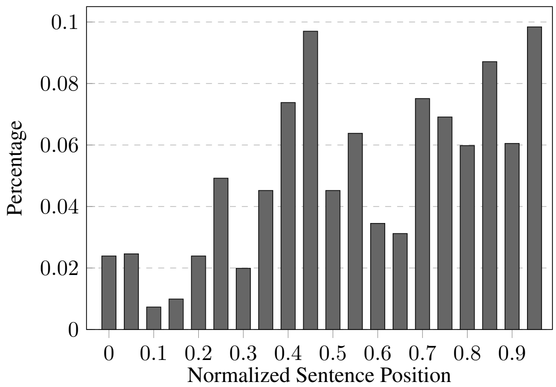


Figure 2: The distribution of rationales over sentences of the introduction section.

Figure 2 shows the distribution of rationales over sentences in the introduction section in a sentence-level view. In such a view, we compute the rationale distribution by normalizing each sentence position concerning the total number of sentences in the introduction section. Rationales are mostly from the middle (35%-55%) and last (70%-100%) sentences of the introduction sections. The middle part of the introduction sections describes the research method used in the corresponding scientific paper. The last part reports experiments and findings. We conclude that dialogues explore the scientific contributions of a scientific paper from different scientific views.

EXP and ARG Intents. To study to what degree the dialogues are exploratory and argumentative, we instruct four NLP experts to annotate all the sentences from ArgSciChat dialogues. We remark that these experts are not from those that participated in our dialogue collection study. Each expert annotates 25% of ArgSciChat sentences. As instruction, we give them our intent definitions and their example sentences (Table 1 from Section 3). The average Fleiss' Kappa (Fleiss, 1971) is 0.89, and the Krippendorf's alpha (Krippendorff, 2019) is 0.83, showing that the quality of annotations is high enough to be used in our analysis of the EXP

[4]We use `all-mpnet-base-v2` as the current best performing model for sentence representation.

Role	AI	ST	GI	GO	AO	EXP	ARG
P	63.3	1.2	7.6	21.9	6.1	72.0	28.0
E	1.0	7.6	72.7	11.8	6.9	81.3	18.7
Total	29.5	4.7	42.9	16.4	6.5	77.1	22.9

Table 4: The frequency (%) of intents from each dialogue role. "Total" is the total frequency of each intent.

Model	Rationale-F1	Message-F1	BScore	Mover
TF-IDF	16.22	-	-	-
S-BERT	13.65	-	-	-
LED(Q,P)	10.58	14.25	85.85	2.25
LED(Q,P,H)	8.50	16.14	86.00	4.54
LED(Q,R)	-	19.54	86.64	8.53
Human	51.26	-	-	-

Table 5: LED's scores for rationale selection (Rationale-F1) and response generation (Message-F1, BScore, and Mover).

Query
Sounds promising! So how well do the models perform?

LED(Q,P):
We use a neural network to train the neural network to generate neural networks.
LED(Q,P,H):
We use a generic, KB-agnostic annotation approach.
LED(Q,R):
They use their models to improve annotation speed. What do you think about this?

Reference answer:
Using these in a user study, we measured an annotation time improvement of up 35%. Do you think this is useful?

Reference rationale:
We show that statistical machine learning models can be used in an interactive entity linking setting to improve annotation speed by over 35%.

Table 6: A case study of the response generated by the LED agent when trained on the ArgSciChat dataset.

and ARG intents of dialogues in ArgSciChat.

Table 4 reports the frequency of each intent label on all the sentences expressed by P and E from all dialogues in ArgSciChat. P is more inclined to ask questions (AI), whereas E reports about rationales from paper content (GI). This is compatible with our definitions for these roles, showing that our dialogue formulation was comprehensive for the expert subjects to follow. P expresses an opinion (GO) more often than E by 10%. Both P and E ask for opinion (AO) with a similar frequency (≈6%). E switches the topic of conversation (ST) more frequently than P by about 6%, which is because E has access to more content of the scientific papers than P. According to our formulation, E should drive the conversation to prevent any dead end.

By grouping intent labels according to their corresponding intent categories (the EXP and ARG columns in Table 4), we observe that ARG intents represent nearly 23% of all the sentences. In comparison, EXP intents represent about 77% of interactions. These results support the correctness of our dialogue collection methodology in achieving the goal of collecting dialogues that demonstrate EXP and ARG dialogues in a scientific domain.

5.3 Using ArgSciChat to Evaluate Dialogue Agents

Table 5 shows the results for both rationale selection and answer generation tasks (Section 4). For rationale selection, the performance of LED in terms of Rationale-F1 is lower than the baseline retrieval-based methods and far from human scores. This confirms the difficulty of ArgSciChat for the rationale selection task, which could happen because of the small number of data samples. This motivates further research in developing few-shot learning methods for training conversational agents in specialized domains.

For response generation, in terms of the Message-F1 metric, LED(Q, P) achieves a 14.25% token-level F1 score, followed by the LED(Q,P,H) with a 16.14% score. LED(Q,P,H) outperforms LED(Q,P), suggesting that prior exchanged messages in a dialogue can be beneficial in generating a response to an input query. This also shows that dialogues in our dataset are beyond a sequence of unrelated dialogue turns. LED(Q,R) outperforms LED(Q,P) and LED(Q,P,H) by at least three token-level F1 points. We observe a similar trend for BScore and Mover, confirming the importance of rationales for generating human-like answers.

As a case study, we analyze responses LED generates for a query given different inputs. Table 6 shows the query, the responses generated by the examined configurations of LED, the response generated by E, and the rationales selected by E in a dialogue. LED(Q,R) identifies Q's intent, which aims to ask for information about "the speed improvement of the annotation study". The model generates a response with both EXP and ARG intents. The first sentence in the LED(Q,R)'s response has EXP intent as it provides an informative response. The second sentence involves argumentation by asking for P's opinion, heading towards an argumentative dialogue. Although a few training examples are given, they seem to be curated to grasp the scientific argument generation.

6 Conclusions

We propose a scientific argumentative dialogue dataset in the NLP scientific domain. In our dataset, dialogue sentences are in free-form English texts, written by NLP experts, and annotated with intent labels encoding the exploratory and argumentative intents of scientific argumentation. To collect this dataset, we define a new methodology, letting scientists introduce their scientific papers, choose the topic and the time of dialogues based on their preferences. In addition to in-depth analysis of our datasets, we report it as challenge for a strong dialogue agent. As shown in our experiments, even a small-sized but high-quality dataset contributes to developing dialogue agents in specialized domains.

Ethics Statement

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.

Limitations

ArgSciChat is the result of a pilot study concerning 31 invited NLP experts. In particular, ArgSciChat contains dialogues about 20 scientific papers regarding a few NLP topics. Thus, dialogues in ArgSciChat are only a small sample of the set of possible dialogues grounded in scientific papers. In particular, several design choices have been considered in our data collection methodology: (a) the topic of a paper; (b) the common background of invited NLP experts; (c) the available content of a paper during a dialogue; and (d) the dialogue setting (e.g., in our implementation, dialogues had a time limit which restricted the number of interactions between subjects). We chose the NLP domain in our study since we (the authors) have expertise in this domain. This choice also facilitated the definition of a pool of NLP experts to participate in our study through our research network.

Furthermore, dialogues in ArgSciChat are grounded in scientific papers. In particular, we limit the paper's content to the abstract and introduction sections. This choice reduces subjects' effort to act as `E`, while also providing enough information to sustain a dialogue. Thus, we do not collect dialogues between subjects concerning other sections of a paper.

Acknowledgments

This work has been funded by the LOEWE Distinguished Chair "Ubiquitous Knowledge Processing" (LOEWE initiative, Hesse, Germany), by the German Research Foundation (DFG) as part of the QASciInf project (grant GU 798/18-3), and by the European Union's Horizon 2020 research and innovation program under grant agreement 101017142.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. 2020. APE: Argument pair extraction from peer review and rebuttal via multi-task learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7000–7011, Online. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online. Association for Computational Linguistics.

Bettina Fazzinga, Andrea Galassi, and Paolo Torroni. 2021. An argumentative dialogue system for COVID-19 vaccine information. In *Logic and Argumentation - 4th International Conference, CLAR 2021, Hangzhou, China, October 20-22, 2021, Proceedings*, volume 13040 of *Lecture Notes in Computer Science*, pages 477–485. Springer.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 8118–8128, Online. Association for Computational Linguistics.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

G. Nigel Gilbert. 1976. The transformation of research findings into scientific knowledge. *Social Studies of Science*, 6(3/4):281–306.

G. Nigel Gilbert. 1977. Referencing as persuasion. *Social Studies of Science*, 7(1):113–122.

Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. 2019. Argument mining for understanding peer reviews. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2131–2137, Minneapolis, Minnesota. Association for Computational Linguistics.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1647–1661, New Orleans, Louisiana. Association for Computational Linguistics.

Martin Krallinger, Anastasia Krithara, Anastasios Nentidis, Georgios Paliouras, and Marta Villegas. 2020. BioASQ at CLEF2020: Large-scale biomedical semantic indexing and question answering. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR, Lisbon, Portugal*, page 550–556, Berlin, Heidelberg. Springer-Verlag.

Klaus Krippendorff. 2019. Content analysis: An introduction to its methodology.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Rafael Mestre, Razvan Milicin, Stuart E. Middleton, Matt Ryan, Jiatong Zhu, and Timothy J. Norman. 2021. M-arg: Multimodal argument mining dataset for political debates with audio and transcripts. In *Proceedings of the 8th Workshop on Argument Mining*, pages 78–88, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Randall Munroe. 2013. The rise of open access. *Science*, 342(6154):58–59.

Sheshera Mysore, Arman Cohan, and Tom Hope. 2022. Multi-vector models with textual guidance for fine-grained scientific document similarity. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4453–4470, Seattle, United States. Association for Computational Linguistics.

Matan Orbach, Yonatan Bilu, Ariel Gera, Yoav Kantor, Lena Dankin, Tamar Lavee, Lili Kotlerman, Shachar Mirkin, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. A dataset of general-purpose rebuttal. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5591–5601, Hong Kong, China. Association for Computational Linguistics.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.

Dimitris Pappas, Ion Androutsopoulos, and Haris Papageorgiou. 2018. BioRead: A new dataset for biomedical reading comprehension. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dimitris Pappas, Petros Stavropoulos, Ion Androutsopoulos, and Ryan McDonald. 2020. BioMRC: A dataset for biomedical machine reading comprehension. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 140–149, Online. Association for Computational Linguistics.

Daraksha Parveen, Mohsen Mesgar, and Michael Strube. 2016. Generating coherent summaries of scientific articles using coherence patterns. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 772–783, Austin, Texas. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Francesco Ronzano and Horacio Saggion. 2015. Dr. inventor framework: Extracting structured information from scientific publications. In *Discovery Science - 18th International Conference*, volume 9356 of *Lecture Notes in Computer Science*, pages 209–220, Banff, AB, Canada.

Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2018. A survey of available corpora for building data-driven dialogue systems: The journal version. *Dialogue Discourse*, 9(1):1–49.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, Liat Ein-Dor, Roni Friedman-Melamed, Assaf Gavron, Ariel Gera, Martin Gleize, Shai Gretz, Dan Gutfreund, Alon Halfon, Daniel Hershcovich, Ron Hoory, Yufang Hou, Shay Hummel, Michal Jacovi, Charles Jochim, Yoav Kantor, Yoav Katz, David Konopnicki, Zvi Kons, Lili Kotlerman, Dalia Krieger, Dan Lahav, Tamar Lavee, Ran Levy, Naftali Liberman, Yosi Mass, Amir Menczel, Shachar Mirkin, Guy Moshkowich, Shila Ofek-Koifman, Matan Orbach, Ella Rabinovich, Ruty Rinott, Slava Shechtman, Dafna Sheinwald, Eyal Shnarch, Ilya Shnayderman, Aya Soffer, Artem Spector, Benjamin Sznajder, Assaf Toledo, Orith Toledo-Ronen, Elad Venezian, and Ranit Aharonov. 2021. An autonomous debating system. *Nature*, 591(7850):379–384.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artières, Axel-Cyrille Ngonga Ngomo, Norman Heino, Éric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138:1–138:28.

David Wadden, Kyle Lo, Lucy Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 61–76, Seattle, United States. Association for Computational Linguistics.

Douglas Walton. 2008. *Informal Logic: A Pragmatic Approach*, 2 edition. Cambridge University Press.

Douglas Walton and Fabrizio Macagno. 2007. Types of dialogue, dialectical relevance and textual congruity. *Anthropology and Philosophy*, 8(1-2):101–120.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 129–136, Sapporo, Japan. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR*, Addis Ababa, Ethiopia. OpenReview.net.

Zhiwei Zhang, Jiyi Li, Fumiyo Fukumoto, and Yanming Ye. 2021. Abstract, rationale, stance: A joint model for scientific claim verification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3580–3586, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

A Dialogue Collection Tool Supplementary

We provide details about our implementation of the described data collection methodology below. We define a time deadline for each step to ease subjects' synchronization along the data collection process.

We require subjects to provide the following contact information: full name and email address. We also need subjects to provide their Google Scholar profiles regarding paper selection. If no Google Scholar profile is available, subjects can still sign-up by manually submitting hyperlinks to the PDF files of papers about which they have enough expertise. The automatic paper retrieval step reports the top five subjects' most recent papers listed in their Google Scholar profile. Subjects are then asked to select two of them to participate in the study.

We consider time slots of one hour in length. Upon sign-up, summary information about selected papers and time slots is provided to human subjects. In particular, subjects can add the selected time slots to their calendar to ensure participation.

Once the sign-up deadline has been met, our implementation automatically notifies subjects about the next phase via email. Additionally, a unique private authentication code is assigned and provided to each participant. The authentication code is used to identify subjects during the study while ensuring anonymity. We split each `E` time slot into three time slots of 20 minutes duration. Subjects are required to book four distinct time slots, each one regarding a different paper. Upon submission, summary information is provided to subjects regarding their final schedule concerning both roles.

Once the previous step time deadline has been met, our implementation automatically notifies subjects about their final schedule. By doing so, subjects have exact information about which time slots they have to join. Our implementation also supports an automatic notification system that ensures subjects join each time slot on time. Lastly, during a dialogue, we do not enforce any explicit intent on participants' dialogue turns to avoid potential biases in collected data.

Why Synchronous Dialogues? Given subjects' required level of expertise and to ensure naturalness during dialogue collection, we opted for synchronous dialogues. Indeed, such a choice introduces tight requirements about subjects' availability and corresponding attributable effort. However, asynchronous dialogues can also have potential drawbacks, such as an extended data collection period. Additionally, in an asynchronous dialogue, dialogue properties like the partial observability of `P` can be quickly overcome due to the absence of a time limit. Conversely, short synchronous dialogues encourage both subjects to maximize the exchanged information flow in the same way as a natural goal-oriented conversation.

Data Collection Interfaces We developed role-specific collection interfaces based on the Mephisto library[5]. Upon connection of both subjects for a dialogue, the corresponding paper content is loaded, and dialogue roles are attributed to each subject. To avoid a `E` message without rationales, the system notifies when no rationales have been highlighted upon a message sent. Additionally, a hint system encourages both subjects to suggest discussing previous dialogue turns, encouraging argumentative messages. Figure 3 reports the data collection interfaces for `P` and `E`.

B Dialogue Diversity Supplementary

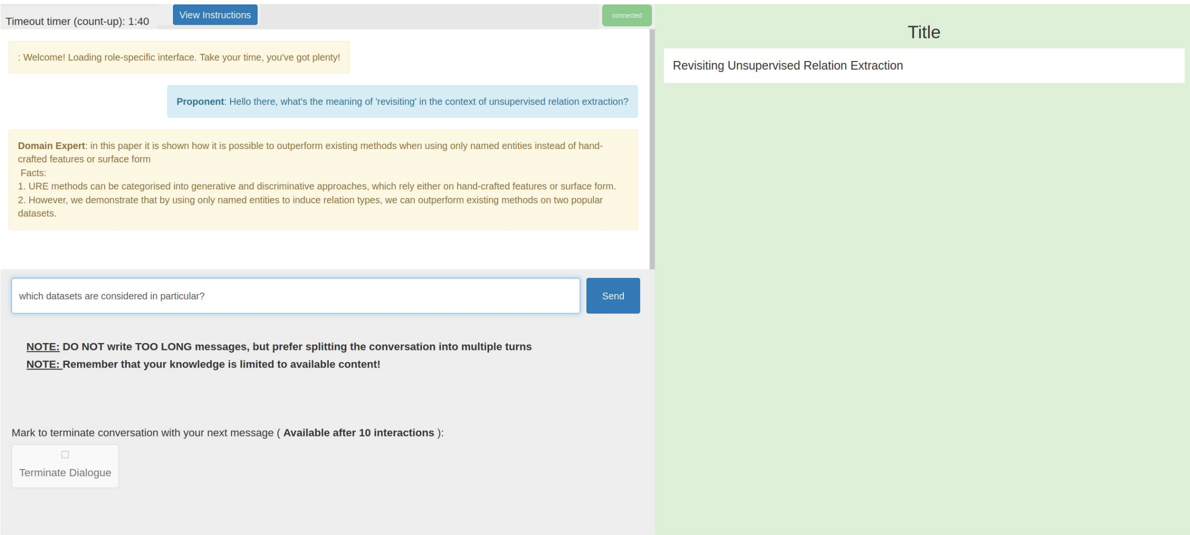
Table 7 shows the diversity scores for each group of messages across different diversity threshold values and sentence transformer models. In particular, we consider one of the best performing general-purpose sentence transformers (*all-mpnet-base-v2*) and the sentence transformer versions of PubMedBERT finetuned on SciFact (*S-PubMedBert-MS-MARCO-SCIFACT*) and SciBERT (*S-Scibert-snli-multinli-stsb*). We observe similar trends concerning semantic diversity across groups in all employed models, highlighting the quality of collected data.

C Using ArgSciChat to Evaluate Dialogue Agents Supplementary

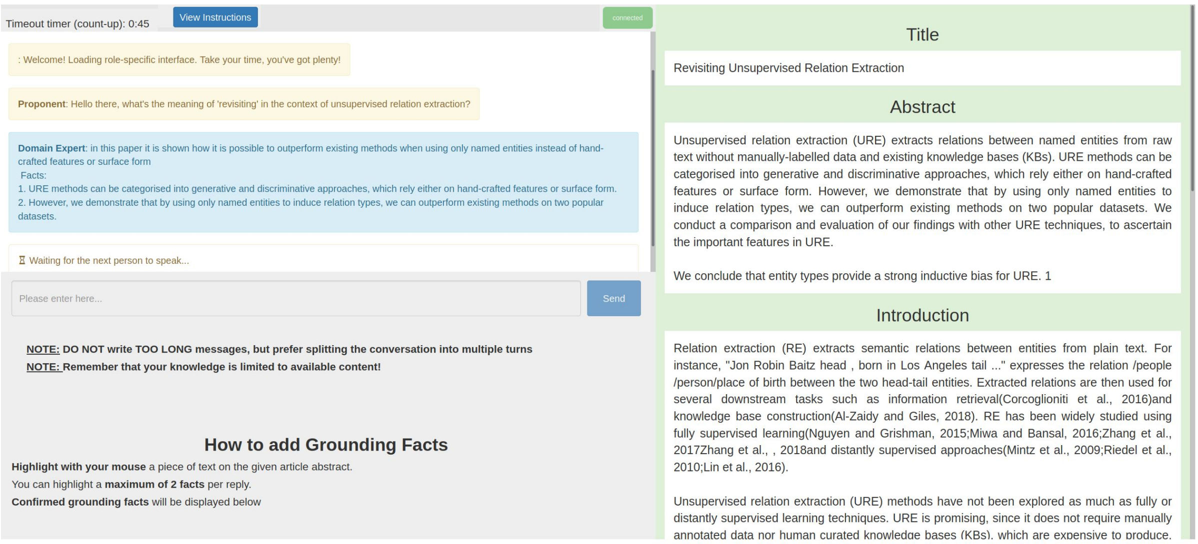
In this section, we report additional details regarding our experimental setup and the described tasks.

Setup As in Dasigi et al. (2021), multiple inputs are concatenated together with a special separation token. In our experimental setup, inputs are concatenated based on the following order: (i) query (Q); (ii) dialogue history (H); (iii) scientific paper (P). We set the LED global attention mask to take into account each individual input as in (Dasigi et al., 2021). In contrast to (Dasigi et al., 2021), the

[5]https://github.com/facebookresearch/Mephisto



(a)



(b)

Figure 3: Data collection interfaces for (a) P and (b) E. On the right side, the title, abstract and introduction sections of the paper are reported for E. Conversely, P's view is solely restricted to the paper title.

lack of LaTeX source files forbids the automatic retrieval of paper section paragraphs. Instead, we work at the sentence level. The average training time is $\sim$40 $-$ 60 seconds per epoch depending on the given input combination. Experiments are conducted on a NVidia GeForce 2080ti 11 GB. Lastly, Table 8 shows fold splits statistics.

Model Configuration Table 9 reports the main hyper-parameters of the LED model. As in (Dasigi et al., 2021), we employ the HuggingFace model `allenai/led-base-16384`. Because we only use the abstract and introduction sections of a scientific paper, we re-scale the *attention window size* hyper-parameter according to our maximum document length. Note that the selected hyper-parameter value is the same percentage of the maximum document length as in (Dasigi et al., 2021).

Group	0.3	0.4	0.5	0.6	0.7
`S-BERT`					
G1	0.77	0.87	0.92	0.96	0.98
G2	0.43	0.48	0.63	0.73	0.68
G3	0.18	0.28	0.43	0.60	0.81
`SciFact`					
G1	0.61	0.82	0.90	0.95	0.97
G2	0.41	0.53	0.63	0.73	0.79
G3	0.11	0.21	0.35	0.49	0.70
`SciBERT`					
G1	0.37	0.67	0.86	0.93	0.97
G2	0.48	0.68	0.75	0.82	0.90
G3	0.10	0.20	0.36	0.56	0.77

Table 7: Average topical diversity on dialogues on identical papers. The numbers in the header of columns represent threshold values.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Train	199	200	183	212	207
Validation	27	14	39	27	22
Test	23	35	27	10	20

Table 8: The number of dialogue turns in each fold for training and evaluating LED agent on our dataset.

Hyperparameter	Value
Attention dropout	0.1
Attention window size	700
Optimizer	Adam
Learning rate	$5 \cdot 10^{-5}$
Patience	20

Table 9: The LED model hyper-parameters.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
8

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

1

☐ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
1

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Appendix

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
4

C ☑ Did you run computational experiments?

4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
4

D ☑ **Did you use human annotators (e.g., crowdworkers) or research with human participants?**
Left blank.

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Appendix

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
3