PaperPersiChat: Scientific Paper Discussion Chatbot using Transformers and Discourse Flow Management

Alexander Chernyavskiy
HSE University
Sberbank
Moscow, Russia
alschernyavskiy@gmail.com

Max Bregeda
Moscow State University
Sberbank
Moscow, Russia
mbregeda@gmail.com

Maria Nikiforova
HSE University
Sberbank
Moscow, Russia
labenzom@gmail.com

Abstract

The rate of scientific publications is increasing exponentially, necessitating a significant investment of time in order to read and comprehend the most important articles. While ancillary services exist to facilitate this process, they are typically closed-model and paid services or have limited capabilities. In this paper, we present *PaperPersiChat*, an open chatbot-system designed for the discussion of scientific papers. This system supports summarization and question-answering modes within a single end-to-end chatbot pipeline, which is guided by discourse analysis. To expedite the development of similar systems, we also release the gathered dataset, which has no publicly available analogues.

1 Introduction

Scientific papers are a crucial part of academic research and are used to disseminate new findings, theories and knowledge to the wider community. At the same time, rapid scientific progress makes it challenging to keep up with new technologies without spending a lot of time reading papers. While traditional summarizing services like Elicit[1] and Scholarcy[2] can be helpful, they often unable to explain sophisticated and complex concepts. More advanced solutions, such as Explainthepaper[3], have emerged to address this limitation as they can elucidate user-highlighted text, but also require the user to read the article beforehand.

Dialogue systems are an alternative capable of combining extractive and generative approaches. Grounding-based approaches were suggested to eliminate issues associated with the hallucinations of LMs (Cai et al., 2022; Gao et al., 2022). The release of ChatGPT[4] has propelled chatbots to the forefront of text data processing. The ChatGPT API and proprietary solutions have enabled the creation of communication services like ChatPDF[5] and xMagic[6]. However, these are services with a closed architecture and paid for.

Interestingly, there are no publicly available open systems that do not use the API of LLMs. One of the reasons, is the lack of open-source datasets for dialogue on scientific grounding. To bridge this gap, we present *PaperPersiChat*[7], a chatbot pipeline designed for the scientific paper domain. It capable of communicating on the basis of a user-selected paper by providing summaries and answering clarifying questions. Our second contribution is the training dataset that can be used to develop solutions for similar tasks. Our code is available at https://github.com/ai-forever/paper_persi_chat.

2 Related Work

The incorporation of external information, referred to as grounding, has been shown to enhance the quality of the generation by improving the factual component. Several approaches utilize knowledge bases or web mining (Glaese et al., 2022; Thoppilan et al., 2022), while others focus on extracting information from individual documents. Cai et al. (2022) proposed a transformer-based model which retains context semantics while sacrificing text details due to the use of averaging word embeddings. UniGDD (Gao et al., 2022) and DIALKI (Wu et al., 2021) systems also consider document-grounded generation but are limited by context length or investigated for task formulations different from ours.

The main limitation of such systems is the lack of training datasets. CMU DoG (Zhou et al., 2018) was proposed for grounding-based movie conversations but contains few documents which com-

[1] https://elicit.org
[2] https://www.scholarcy.com
[3] https://www.explainpaper.com
[4] https://chat.openai.com/
[5] https://www.chatpdf.com
[6] https://www.xmagic.ai
[7] *PaperPersiChat* is running online on http://www.PaperPersiChat.tech

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 584–587
September 11–15, 2023.

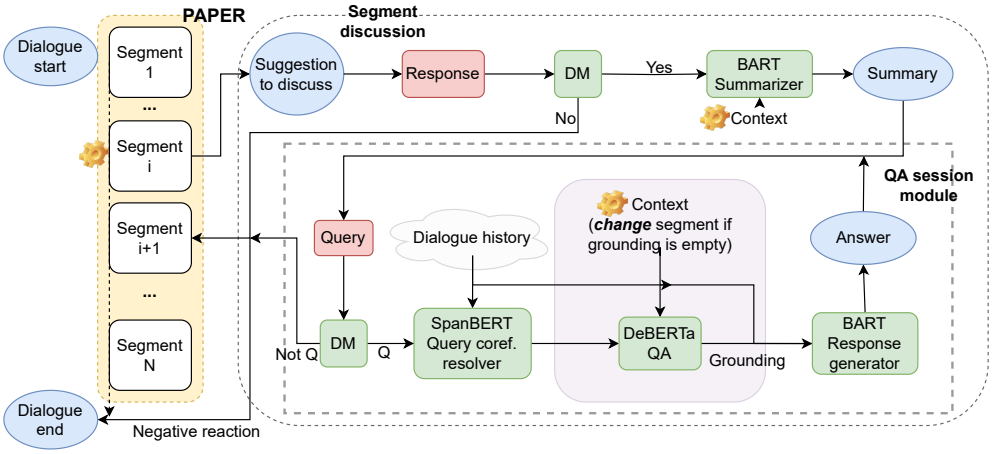

Figure 1: The architecture of *PaperPersiChat*. The discussion of the single i-th segment is demonstrated. User input is shown in red frames, chatbot answers in blue and the trainable pipeline submodules in green. DM refers to the Dialogue Management submodule. The light-purple part of the Question Answering (QA) system runs in a loop over segments while the retrieved grounding is empty. The current segment texts are labelled with a gear icon.

plicates generalization. Larger datasets, such as the Wizard of Wikipedia (Dinan et al., 2019), are labeled more roughly and are not tailored to the scientific domain either. At the same time, recent approaches employ synthetic training datasets collected via ChatGPT (Askari et al., 2023). We follow this idea and propose a dataset collection process, which we used to train our pipeline later.

3 System Overview

Figure 1 shows the general architecture of the *PaperPersiChat* system. The chatbot discusses paper segments step by step, with each segment containing one or several sections of the paper. The dialogue ends when all segments have been discussed or too much negative feedback has been received.

For each segment, the chatbot firstly suggests discussing it. If the suggestion is accepted, it provides a short summary and proceeds to the question-answering session. Otherwise, the chatbot moves to the discussion of the succeeding segment. For each question query, the QA module attempts to extract grounding from the current segment. However, if this fails, it continues to look over all other segments. In case when QA module can't find an answer in the entire paper, it informs the user about that. If the user's query is not a question, the system moves to the discussion of the following segment. Further details are described in Section 5.

4 Data

There is a lack of publicly available datasets for training the dialogue systems with scientific text grounding. Since manual markup requires significant resources, we constructed the dataset automatically. As the source, we used 63,321 computer science papers from the Semantic Scholar Open Research Corpus published at top science conferences between 2000 and 2021. We utilized its subset to collect our dataset, which consists of two parts: instances collected via OpenAI's Davinci or ChatGPT [8].

The Davinci model processed complex instructions and tried to produce the part of the dialogue related to the whole segment discussion part (see Figure 1). In this way, we collected 3,588 raw outputs and each of them was processed further into a summary and dialogue turns. All these summaries were used to train the summarization submodule. Further filtering was done to remove unparsed outputs, short dialogues and dialogues with inconsistent structure (including incorrect speaker order). This yielded a set of 2,817 dialogues that were used to train the models from the QA session module. To construct qualitative dialogues for QA, and also to manage the inputs of the dialogue participants, we used two ChatGPT models talking to each other. The resulting dataset totals 2,817 dialogues produced by Davinci and 8,787 dialogues produced by ChatGPT, with an average of four turns per dialogue. We have made this dataset publicly available via https://huggingface.co/datasets/ai-forever/paper_persi_chat.

[8] https://platform.openai.com/docs/models

5 Submodule Details

This section provides details about submodules of the *PaperPersiChat* pipeline.

Dialogue Discourse Flow Management (DM) This component is employed to classify the user's reaction and navigate to the pertinent pipeline steps. It is composed of two models: a dialogue discourse parser and an agreement classifier. To acquire the discourse parser, we trained the parser proposed by Shi and Huang (2019) from scratch on CDSC (Zhang et al., 2017). To classify the last relation in the dialogue, the pipeline passes the last ten utterances of the dialogue history as the parser input. In this pipeline, we consider only the following dialogues acts: Agreement, Disagreement, Question and Negative Reaction. Since the discourse parser has broad classes and can not distinguish well "yes" and "no" cases, an auxiliary agreement classifier based on the SBERT model (Reimers and Gurevych, 2019) was trained using the open-source CDSC and IAC (Walker et al., 2012) corpora.

Summarization This component briefly summarizes the current segment text. Here, the BART model (Lewis et al., 2020) was selected for generation due to its state-of-the-art performance in similar tasks. We trained it using the 3,204 samples and started from DistilBART[9]. Evaluation of the model on 356 test samples from our dataset yielded ROUGE-1 score of 51.43, ROUGE-2 score of 30.80, and ROUGE-L score of 40.97.

Coreferences Resolver We utilized the pretrained SpanBERT from AllenNLP Framework [10]. Here, we use the concatenation of the user query and four last utterances from the dialogue history as the SpanBERT input and replace entities with coreferences in the query to get the final QA input.

Extractive QA The QA module extracts the most relevant text sufficient to answer the user's question, namely grounding. To ensure the class balance as in SQuAD2.0, we added more hard negative examples (questions with empty answers) by selecting questions generated for similar segments. The similarity scores were calculated as the cosine distance between Specter [11] embeddings. As the final model, we fine-tuned DeBERTA-v3 [12].

Response Generation This component generates the target response text based on the query, dialogue history and grounding text extracted by DeBERTa. We conducted experiments for BART (Lewis et al., 2020) and DialoGPT (Zhang et al., 2020) for two options of groundings: extracted by the pretrained or by the fine-tuned DeBERTa.

To construct model inputs, we concatenated query, dialogue history and grounding via special separation tokens.The BART model trained using groundings from the fine-tuned DeBERTa yielded the best results, with a ROUGE-1 of 61.71 and a BLEU-1 of 50.3 on our test set. In comparison, the BART model trained using groundings extracted by the pretrained QA model got ROUGE-1 of 49.41 and the best DialoGPT model got 61.42.

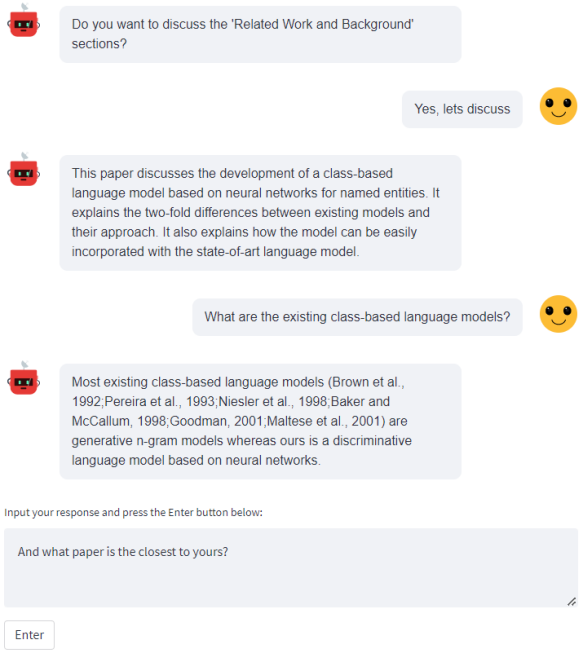

Figure 2: Screenshot of the interface window with a dialogue example generated using *PaperPersiChat*.

6 User Interface

Figure 2 depicts a screenshot of a sample dialogue between the user and the proposed chatbot. Here, the bot suggests discussing the Related Work section; the user agrees and the system moves to the QA session. If during the session the bot cannot find a grounding for a question, it informs the user that there is not enough information in the paper. The QA session continues until the user ceases asking questions, after which the dialogue advances to the next section.

During the dialogue, the user enters his message in the corresponding field and then the dialogue

[9] https://huggingface.co/sshleifer/distilbart-cnn-12-6

[10] https://demo.allennlp.org/coreference-resolution/coreference-resolution

[11] https://huggingface.co/allenai/specter

[12] https://huggingface.co/deepset/deberta-v3-base-squad2

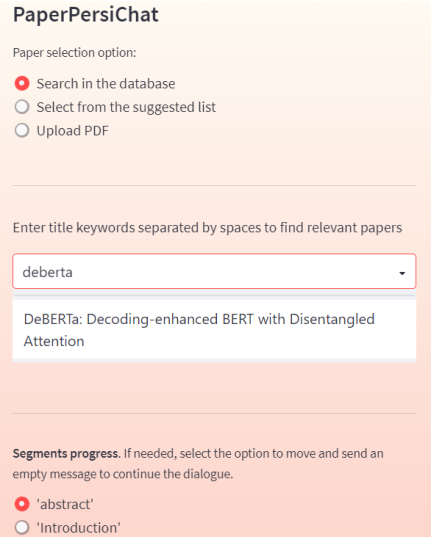

Figure 3: Screenshot of the chat settings.

history above the input field is updated with the addition of the last user's query and the bot's response. Then the process repeats.

The auxiliary menu to the left, illustrated in Figure 3), assists the user in selecting a paper for discussion, switching to another segment (via radio buttons), or clearing the dialogue history. Additionally, the menu provides short instructions to facilitate communication with the bot. There are several options to select a paper for discussion:

- Select any paper from our dataset (63,321 papers) by searching. For this option, the user just needs to enter a few keywords separated by a space and press the "Search" button.
- Select a paper from a suggested sublist.
- Upload new paper in the PDF format.

7 Conclusion

We have presented *PaperPersiChat*, chatbot based only on open-source models and capable of engaging in conversations about scientific papers. For each paper segment, the bot offers the user an opportunity to get a summary and moves to the QA session mode in the case of agreement. The dialogue flow is controlled by a discourse analyzer. We also presented a novel dataset to facilitate the development of similar systems. Future work includes refining individual submodules and dialogue management to promote greater flexibility.

References

Arian Askari, Mohammad Aliannejadi, Evangelos Kanoulas, and Suzan Verberne. 2023. Generating synthetic documents for cross-encoder re-rankers: A comparative study of chatgpt and human experts. *CoRR*, abs/2305.02320.

Yuanyuan Cai, Min Zuo, and Haitao Xiong. 2022. Modeling hierarchical attention interaction between contexts and triple-channel encoding networks for document-grounded dialog generation. In *Frontiers of Physics*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *ICLR 2019*. OpenReview.net.

Chang Gao, Wenxuan Zhang, and Wai Lam. 2022. Unigdd: A unified generative framework for goal-oriented document-grounded dialogue. In *Proceedings of ACL 2022*.

Amelia Glaese et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *CoRR*, abs/2209.14375.

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 ACL 2020*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of EMNLP-IJCNLP 2019*, pages 3980–3990.

Zhouxing Shi and Minlie Huang. 2019. A deep sequential model for discourse parsing on multi-party dialogues. In *Processings of AAAI, 2019*.

Romal Thoppilan et al. 2022. Lamda: Language models for dialog applications. *CoRR*, abs/2201.08239.

Marilyn A. Walker, Jean E. Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of LREC 2012*, pages 812–817.

Zeqiu Wu, Bo-Ru Lu, Hannaneh Hajishirzi, and Mari Ostendorf. 2021. DIALKI: knowledge identification in conversational systems through dialogue-document contextualization. In *Proceedings of EMNLP, 2021*, pages 1852–1863.

Amy X Zhang, Bryan Culbertson, and Praveen Paritosh. 2017. Characterizing online discussion using coarse discourse sequences. In *Eleventh International AAAI Conference on Web and Social Media*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of ACL 2020*.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W. Black. 2018. A dataset for document grounded conversations. In *Proceedings of EMNLP, 2018*.