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

Alexander Chernyavskiy[1,2], Max Bregeda[2,3], Maria Nikiforova[1,2]
[1]HSE University, Russia [2]Sber, Russia
[3]Moscow State University, Russia

## 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.

- We present *PaperPersiChat*, an open chatbot-system designed for the discussion of scientific papers.
- The system supports summarization and question-answering modes within a single end-to-end chatbot pipeline, which is guided by discourse analysis.
- We release the gathered dataset, which has no publicly available analogues.

**Demo:** `http://www.PaperPersiChat.tech`
**Code:** `https://github.com/ai-forever/paper_persi_chat`

## Related Work

- **Summarization:**
  (1) Traditional summarizing services like Elicit and Scholarcy often unable to explain sophisticated and complex concepts.
  (2) More advanced solutions, such as Explainthepaper, require the user to read the article beforehand.
- **Dialogue systems:** The ChatGPT API and proprietary solutions have enabled the creation of communication services like ChatPDF and xMagic. However, these are services with a closed architecture and paid for.
- **Datasets:** Existing open-source datasets are small (CMU DoG) or are roughly labeled (Wizard of Wikipedia) and are not tailored to the scientific domain either.

## Dataset

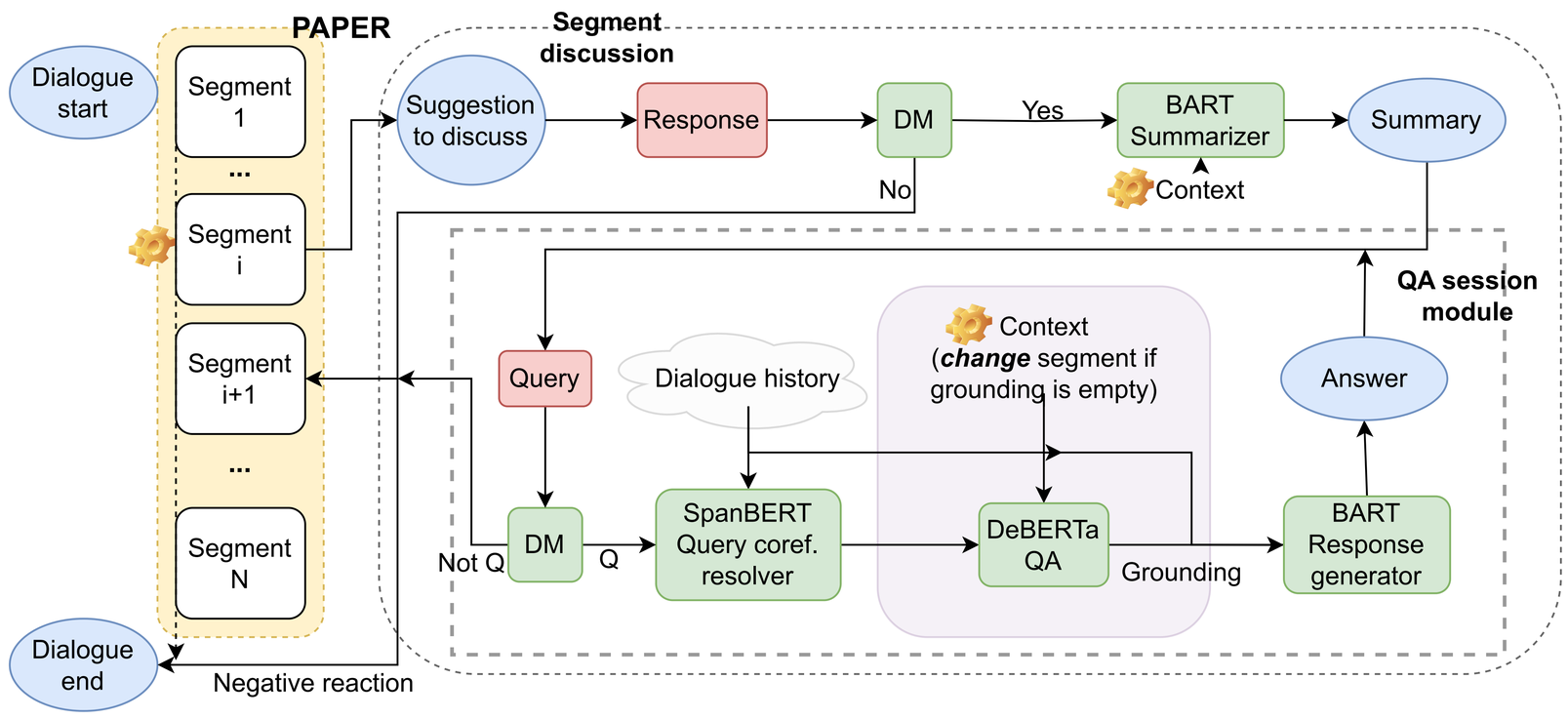
**Source:** 63,321 computer science papers from the Semantic Scholar Open Research Corpus published at top science conferences between 2000 and 2021.

**Parts:**

- Davinci-based
  - ◇ Whole segment discussion
  - ◇ Prompts to generate two outputs: summary and dialogue turns
  - ◇ 2,817 dialogues
- ChatGPT-based:
  - ◇ QA-session dialogues
  - ◇ Two ChatGPT models talking to each other. One of them sees only the summary and asks questions, while the other answers using the full text.
  - ◇ 8,787 dialogues

**Link:** `https://huggingface.co/datasets/ai-forever/paper_persi_chat`

## System Overview

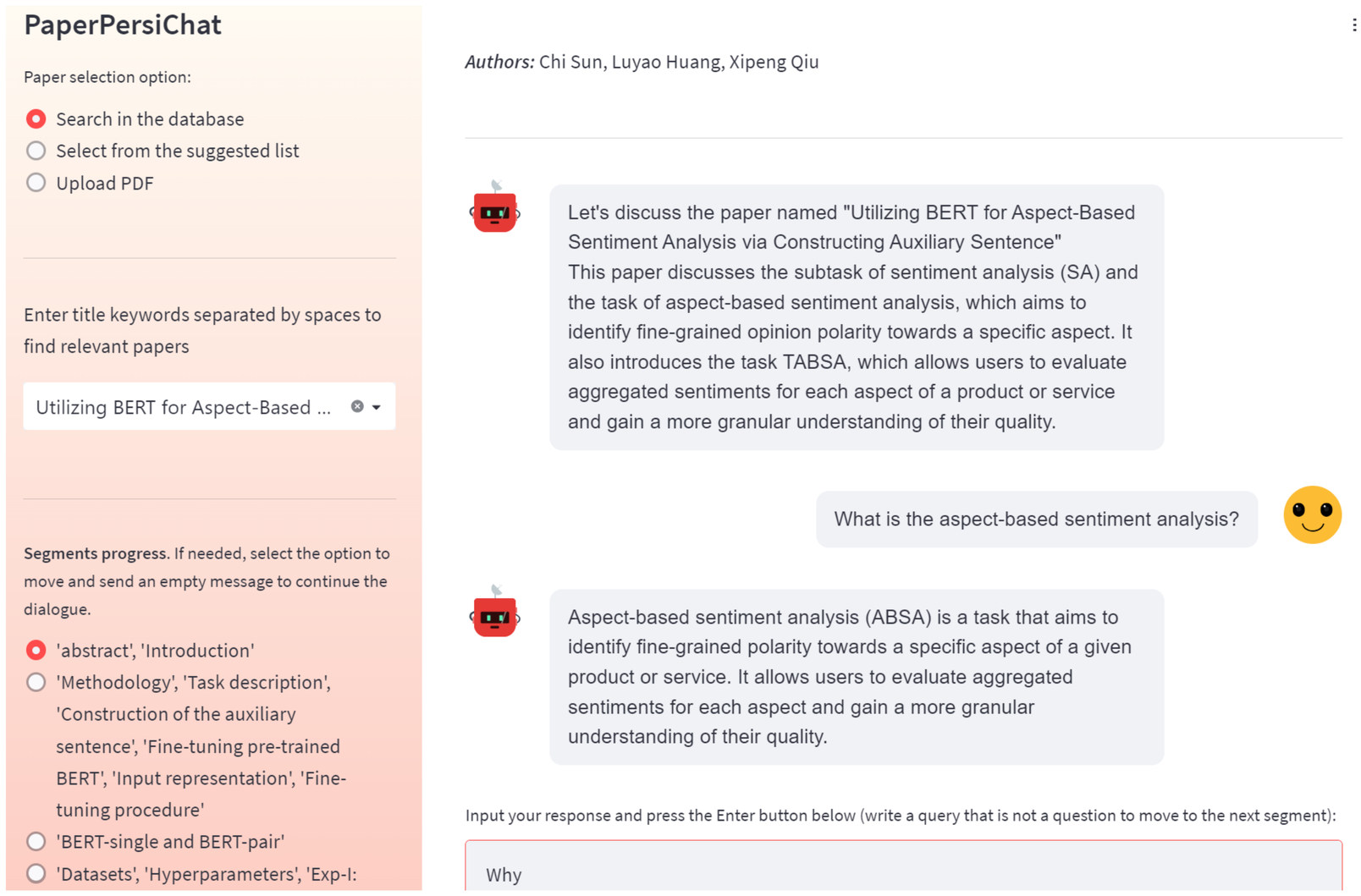



The chatbot discusses paper segments step by step, with each segment containing one or several sections of the paper. The dialogue ends when all segments have been discussed or too much negative feedback has been received.

**Components:**

- ◇ **Dialogue Discourse Flow Management (DM)** Classifies the user's reaction and is used for navigation. Is implemeted via dialogue discourse parser and an agreement classifier (SBERT-based).
- ◇ **Coreferences Resolver** The pretrained SpanBERT from AllenNLP Framework. Input: the concatenation of the user query and four last utterances from the dialogue history. Output: the final QA input.
- ◇ **Extractive QA** Extracts the most relevant text sufficient to answer the user's question, namely grounding.
- ◇ **Response Generation** Generates the target response text based on the query, dialogue history and grounding text extracted by DeBERTa. Input: query, dialogue history and grounding concatenated by separation tokens.

## Interface




**Auxiliary menu:**

- ◇ Paper selection: (1) Select any paper from our dataset by searching; (2) Select a paper from a suggested sublist; (3) Upload new paper in the PDF format.
- ◇ Switching to another segment.
- ◇ Clearing the dialogue history.