WhatsApp Chat Summarization

Andrea Auletta
andrea.auletta@studenti.unipd.it

Davide Baggio
davide.baggio.1@studenti.unipd.it

Marco Bernardi
marco.bernardi.11@studenti.unipd.it

Marco Brigo
marco.brigo@studenti.unipd.it

Sebastiano Sanson
sebastiano.sanson@studenti.unipd.it

Abstract

This research project in Natural Language Processing (NLP) focuses on the development of an automated system for the summarization of chat messages, applicable to both group and individual conversations. The objective is to generate concise summaries of conversations, thereby obviating the need for users to review all messages individually. This application is especially pertinent in contexts where chat logs proliferate quickly, offering significant benefits in terms of time savings and cognitive load reduction for users.

1. Introduction

In the era of digital communication, chat platforms have become fundamentals in our daily interactions, whether they are personal or professional. This new type of transmission favors real-time messages exchange which allows users to engage in conversations that can quickly accumulate extensive chat logs. The huge volume of messages, especially in active group chats formed by dozens if not hundreds of users, often makes it difficult for users to stay updated without dedicating substantial time to review each message. This challenge highlights the need for an efficient method that enables users to quickly extract essential information from chat conversations.

This research project addresses this need by developing an automated system for summarizing chat messages with the goal to generate concise summaries, reducing the need to sort through all messages. By doing so, users can quickly obtain the main points of discussions, significantly saving time and reducing cognitive load.

The significance of this application is evident in various contexts: for instance, in corporate environments, project teams often heavily rely on group chats to coordinate tasks, share updates and exchange miscellaneous informations; similarly, social groups frequently engage in dynamic conversations. In both scenarios, an automated summarization system can enhance productivity and ensure important information is not overlooked.

This research aims to explore and implement state of the art NLP techniques to create an effective summarization system, focusing on the datasets used, the manipulation of them in order to obtain always a coherent data structure, how we obtained new examples to test the performances and the results we got after all.

TO DO: overview of our main results!

2. Datasets

In this section, we explore datasets essential for training and evaluating automated summarization systems for chat messages: an ideal dataset should have a variety of conversation styles and contexts to ensure the robustness and versatility of the summarization models. The datasets should contain real-life or realistic dialogues annotated with concise and accurate summaries. Additionally, the data should reflect diverse communication patterns, including informal and formal registers, and incorporate common conversational elements such as slang, emojis, eventual typos, the use of voice messages and images. We present below two notable datasets, SAMSum and DialogSum (both obtained by Hugging Face repository), which fulfill these criteria and provide rich resources for advancing the field of conversational AI.

2.1. SAMSum

The SAMSum [2] [2] dataset contains about 16,000 messenger-like conversations with summaries splitted in:

- training: 14,732;
- validation: 818;

- test: 819.

Conversations were created and written down by linguists fluent in English. Linguists were asked to create conversations similar to those they write on a daily basis, reflecting the proportion of topics of their real-life messenger conversations. The style and register are diversified - conversations could be informal, semi-formal or formal, they may contain slang words, emoticons and typos. Then, the conversations were annotated with summaries. It was assumed that summaries should be a concise brief of what people talked about in the conversation in third person. The SAMSum dataset was prepared by Samsung R&D Institute Poland and is distributed for research purposes.

2.2. Dialogsum

DialogSum [1] is an extensive dialogue summarization dataset comprising 13,460 dialogues splitted in:

- training: 12,460;
- validation: 500;
- test: 1,500.

supplemented by an additional 100 holdout dialogues designated for topic generation, which contain the id, dialogue, and topic fields. Each dialogue is paired with manually annotated summaries and topics. The dataset is exclusively in English and includes four data fields: the text of the dialogue, a human-written summary of the dialogue, a human-written topic or one-liner of the dialogue, and a unique identifier for each example.

DialogSum distinguishes itself from previous datasets by incorporating dialogues under rich real-life scenarios, including a wider array of task-oriented contexts. The dialogues exhibit clear communication patterns and intents, making them suitable for summarization.

2.3. Custom Dataset

In order to obtain a dataset with an increased volume of data, a broader spectrum of conversation scenarios and an enhanced evaluation, we decided to concatenate the previously mentioned datasets into a bigger one.
The splits of the new dataset are as follows:

- training: 27,192;
- validation: 1,318;
- test: 2,319.

To obtain it we had to preprocess the Dialogsum dataset and... TO DO

2.4. Validation Dataset

To rigorously assess the performance of the selected models within our specific use case, we constructed a validation dataset by generating new examples derived from authentic WhatsApp conversations.

2.4.1 Data Collection

WhatsApp provides a robust functionality for exporting chat conversations in a .txt file format. This exported file encompasses all exchanged messages, complete with timestamps and sender names. Additionally, the file includes various media types, such as images, voice messages, and videos. This comprehensive dataset is instrumental for rigorous evaluation as it reflects real-world conditions and the diverse communication forms encountered in practical applications.

2.4.2 Data Preprocessing

The exported .txt file underwent a meticulous preprocessing phase to extract the pertinent information, including message content and sender details. This information was structured into a Python dictionary as follows:

```
dialogues = [
    {
        'text': "Hello, how are you?",
        'id': 0,
        'golden_summary': "A greeting"
    }
]
```

For conversations containing images or voice messages, the filenames of these media were embedded within the text field to facilitate subsequent processing. As a final step, the filenames were replaced with detailed descriptions for images and transcriptions for voice messages. The models utilized for generating these descriptions and transcriptions are discussed in the following sections.

This preprocessing approach ensures that the dataset is both comprehensive and structured, enabling effective analysis and model evaluation.

3. Models Involved

3.1. BART

TO DO

3.2. Florence-2 Large

To ensure that the images were processed by the summarization model, it was necessary to convert them into detailed textual descriptions. For this task, we employed Florence-2 [4] by Microsoft, an advanced vision foundation

model that utilizes a prompt-based approach to address a wide array of vision and vision-language tasks. Florence-2 is designed to interpret text prompts as task instructions and produce the desired textual outputs, including captioning, object detection, grounding, and segmentation. This model is built upon the extensive FLD-5B dataset, which comprises 5.4 billion annotations across 126 million images, thereby enhancing its multi-task learning capabilities. The sequence-to-sequence architecture of Florence-2 allows it to perform effectively in both zero-shot and fine-tuned scenarios, establishing itself as a robust and competitive vision foundation model.

3.3. Whisper

In addition to processing images, it was necessary to convert voice messages to text. For this task, we utilized Whisper by OpenAI [3], an automatic speech recognition (ASR) system. Whisper is trained on 680,000 hours of multilingual and multitask supervised data collected from the web. The extensive and diverse dataset enhances the system's robustness to accents, background noise, and technical language, which is particularly beneficial for our use case, as voice messages are often of suboptimal quality. Furthermore, Whisper supports transcription in multiple languages and translation from those languages into English.

4. Fine Tuning

5. Testing

6. Results

References

[1] Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. DialogSum: A real-life scenario dialogue summarization dataset. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online, Aug. 2021. Association for Computational Linguistics.

[2] Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *CoRR*, abs/1911.12237, 2019.

[3] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[4] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks, 2023.