GIST: Transforming Overwhelming Information into Structured Knowledge with Large Language Models

Meng Wu, Xinyu Zhou, Gang Ma, Zhangwei Lu, Liuxin Zhang, and Yu Zhang(✉)

Lenovo Research, Beijing, China
Zhangyu29@lenovo.com

Abstract. This paper introduces GIST (Generative Information Synthesis Taskforce), a novel personal knowledge management system that utilizes large-scale online language models to analyze and organize the information, generating structured results, including summaries, key points, and questions and answers. The system also utilizes a multimodal information processing approach to enhance comprehension of the content. As the user's knowledge base grows, GIST becomes a personal knowledge database and provides the necessary information at the right moment. GIST can be accessed on any device, serving as the brain and soul of the user's devices, and empowering them to effectively manage their personal knowledge. Our demo video is at https://youtu.be/ImtduHMQKFQ.

Keywords: Personal Knowledge Database · Multimodal Information Processing with LLM · Structured Knowledge

1 Introduction

The evolution of societal requirements and interactions has undergone significant changes in different eras. The era of computation saw the widespread availability of computational resources, making general computing a commodity that was affordable and accessible to more people. This led to the emergence of task processing for information digitization, which was pioneered by Microsoft. The primary interaction methods during this era were command sets and WIMP (Windows, Icons, Menus, Pointer) [1]. In the internet era, the ubiquitous accessibility of information resources made general information a commodity, leading to services for information acquisition, sharing, and consumption that constitute the entire internet ecosystem. The interaction methods mainly revolved around WIMP, touch, and recommendations, delivering information-related services anytime and anywhere. The interaction principle here is mobile, intuitive, invisible [2]. In the era of large-scale language models (LLM), the universalization of knowledge resources prompted the demand and business model for refined, customized knowledge management. Interaction emphasis was placed on spontaneous interaction and implicit interaction.

As we transition from the internet era to the age of large models, we process a wealth of information daily, ranging from work reports, meetings, books, videos, to our own

L. Fang et al. (Eds.): CICAI 2023, LNAI 14474, pp. 41–45, 2024.
https://doi.org/10.1007/978-981-99-9119-8_4

thoughts and ideas. However, this information is scattered across various devices and apps, making it difficult to find when we need it. For instance, we might bookmark webpages on our computers, like videos on TikTok on our mobile phones, or take notes during an open course on our tablets. Gist utilizes a multimodal information processing method with LLM to consolidate these fragmented pieces of information into structured knowledge. This system can analyze and generate structured results from audio, video, or text data, regardless of the mode of interaction. As the system accumulates knowledge and becomes familiar with the user's cognitive habits, it can provide the necessary knowledge at the right time, even becoming the user's digital alter ego. This assistant can reside in any device the user uses, operating both online and offline, thus creating a private, personalized knowledge base.

2 System Architecture

The system comprises several modules, including a vector database, an Automatic Speech Recognition (ASR) model, a large-scale language model, a module for processing multimodal information, a module for prompt engineering, a module for question-answering flow, and a user interface. The system workflow is presented as follows (Fig. 1).

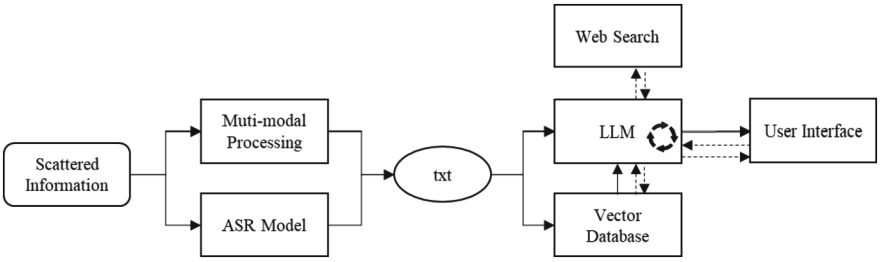

Fig. 1. The workflow of the system is illustrated by the solid line, depicting the automatic generation of summaries and themes by the system. The dotted line represents the flow of user and system interactions in the free chat feature.

A vector database is a specialized database that stores data in the form of vectors, which are mathematical objects possessing both magnitude and direction [3]. In a vector database, each data item is represented as a vector, enabling the storage of information such as position, velocity, or orientation. By encoding personal information as vectors, with the magnitude of each vector conveying the data and the direction providing additional context, the vector database can quickly and easily retrieve information, such as title or author, without requiring searches through traditional databases or spreadsheets. Furthermore, new information can be readily incorporated into the database by simply adding new vectors, making it a suitable method for managing personal collections of movies and books [4].

An ASR model is a type of machine learning model that is used to recognize spoken language in a variety of applications, such as voice assistants, dictation software, and

call centers. The model takes as input an audio recording of speech, and outputs a transcription of the speech in written form. Besides, distinguish between different speakers [5].

A large-scale language model is a type of machine learning model that is used to generate text. LLMs are particularly useful for understanding the meaning of text because they are trained on large amounts of data and can learn to recognize patterns and relationships within language. This allows them to make connections between words and phrases, and to identify the underlying structure and organization of a document, even if it is complex and contains multiple ideas and themes [6]. LLMs provide the summary, theme, key point, Q&A, ToDo list according to the prompt.

Multimodal information processing refers to the ability of a system to process and analyze information that is presented in multiple forms, or modalities. This can include text, images, audio, video, and other types of data. The goal of multimodal information processing is to combine and integrate information from different modalities to gain a more complete and accurate understanding of a given situation or task [7]. For example, a system that uses multimodal information processing might be able to recognize and transcribe spoken words, while also analyzing the facial expressions and body language of the speaker to gain additional insights into their meaning and intent [8].

Prompt engineering is the process of designing and crafting input prompts for machine learning models, with the goal of obtaining the desired output or behavior from the model. A prompt can take many forms, including a natural language utterance, a visual or auditory cue, or a specific set of inputs or parameters. Prompt engineering is an important aspect of our system, as the quality and effectiveness of the prompt can have a significant impact on the accuracy and efficiency of the model's output. To guarantee the quality of the output, the system's prompts must encompass language limits, task requirements, further specific requirements, and the output format and presentation [9].

Question-answering flow refers to the process of determining whether a given question can be answered by a large pre-trained language model or whether additional information needs to be obtained from external sources, such as the internet. The basic idea behind question-answering flow is to use the LLM to determine whether the question can be answered based on the information contained within the original text, or whether additional information needs to be obtained from external sources. The goal of question-answering flow is to provide a seamless and efficient way to obtain answers to questions, whether they can be answered by the model or not. This can help to improve the user experience and enhance the overall effectiveness of the system.

The user interface of the system is designed to preserve the original content of the text and to highlight and structure the knowledge present in the text. To this end, the interface displays the original text or transcript of the conversation, along with a mind map that indexes the main ideas and topics discussed. Users can also access an editable summary of the entire conversation, which allows them to review and modify the machine-generated summary to ensure its accuracy. Additionally, the user interface provides a free dialog box where users can type in their own questions or concerns, which the system will then attempt to address using the structured knowledge it has learned from the conversation. This user-friendly interface allows users to easily navigate and interact with the system, making it a valuable tool for conducting natural language conversations.

3 Demo Procedure and Key Features

In the current stage, we have completed the first step of extracting structured knowledge from information. We have built a demo using online LLMs and ASR models. Users can upload audio or video files, and the system will automatically generate transcripts, perform content analysis, and output summaries, themes, key points, Q&A, and ToDo lists. If the generated content is not satisfactory, users can modify the content by adding or deleting parts and then regenerate it. They can also modify the generated content or trace it back to the original text (Fig. 2).

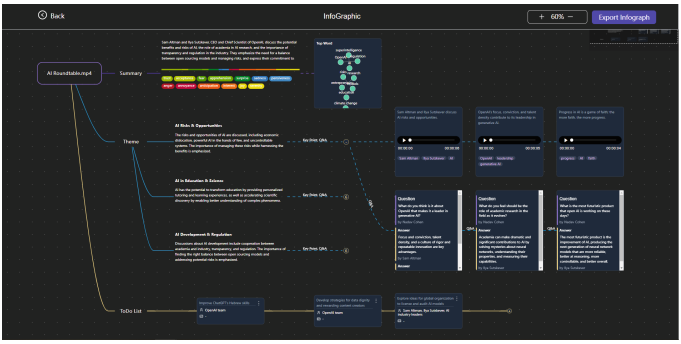

Fig. 2. Analysis user interface include summaries, themes, key points, Q&A, ToDo lists, etc.

To sum up, the key features of GIST are as followed:

1. Efficiently extract structured knowledge from scattered information.
2. Be compatible with multiple devices including PCs, smartphones, tablets, and wearable devices.
3. Support multiple modalities such as text, images, audio, and video.
4. Support multiple scenarios and provide a visualized and interactive experience.

4 Conclusion

The Gist framework successfully organizes fragmented information into structured knowledge, enabling users to browse, understand, and memorize the information while also facilitating direct use of structured knowledge with traceback from knowledge to information. Building upon the successful demonstration of extracting structured knowledge from multimodal/multiple-source fragmented information, we aim to systematically integrate personal and general knowledge, and continue to improve task perception and inference in implicit interaction scenarios based on multimodal information input. Our ultimate goal is to establish a personalized knowledge platform ecosystem that leverages the capabilities of Gist to provide users with a seamless and efficient way to access and use structured knowledge.

References

1. Myers, B.A.: A brief history of human-computer interaction technology. Interactions, **5**(2), 44–54 (1998)
2. Piccolo, L.S.G., De Menezes, E.M., De Campos Buccolo, B.: Developing an accessible interaction model for touch screen mobile devices: preliminary results. Presented at the Proceedings of the 10th Brazilian Symposium on Human Factors in Computing Systems and the 5th Latin American Conference on Human-Computer Interaction, pp. 222–226 (2011)
3. Stata, R., Bharat, K., Maghoul, F.: The term vector database: fast access to indexing terms for web pages. Comput. Netw. **33**(1–6), 247–255 (2000)
4. Lobentanzer, S., Saez-Rodriguez, J.: A platform for the biomedical application of large language models. arXiv preprint arXiv:2305.06488 (2023)
5. Gudepu, P.R., et al.: Whisper augmented end-to-end/hybrid speech recognition system-CycleGAN approach. Presented at the INTERSPEECH, pp. 2302–2306 (2020)
6. OpenAI, "GPT-4 Technical Report." https://cdn.openai.com/papers/gpt-4.pdf. Accessed 28 June 2023
7. Sarter, N.B.: Multimodal information presentation: Design guidance and research challenges. Int. J. Ind. Ergon. **36**(5), 439–445 (2006)
8. Khullar, A., Arora, U.: MAST: multimodal abstractive summarization with trimodal hierarchical attention. arXiv preprint arXiv:2010.08021 (2020)
9. Ekin, S.: Prompt Engineering for ChatGPT: A Quick Guide to Techniques, Tips, and Best Practices (2023). https://doi.org/10.36227/techrxiv.22683919