ConvLogRecaller: A Real-Time Conversational Lifelog Recaller

Yuan-Chi Lee
Department of Computer Science and Information Engineering, National Taiwan University
Taipei, Taiwan
yclee@nlg.csie.ntu.edu.tw

An-Zi Yen
Department of Computer Science, National Yang Ming Chiao Tung University
Hsinchu, Taiwan
azyen@nycu.edu.tw

Hen-Hsen Huang
Institute of Information Science, Academia Sinica
Taipei, Taiwan
hhhuang@iis.sinica.edu.tw

Hsin-Hsi Chen
Department of Computer Science and Information Engineering, National Taiwan University
Taipei, Taiwan
hhchen@ntu.edu.tw

ABSTRACT

The popularization of networks fosters the convenience of communication. People can easily share their life experiences and thoughts with relatives and friends via instant messaging software. As time passes, individuals may forget certain details of life events, leading to difficulties in effectively communicating with others. The propensity of individuals to forget or mix up life events highlights the importance of services aimed at retrieving information about past experiences. This paper presents a conversational information recall system, ConvLogRecaller, which proactively supports real-time memory recall assistance during online conversations. Given a conversation of the user with others, ConvLogRecaller suggests a message if the user forgets the details of the life experiences. The services provided by our system can avoid hesitations or memory lapses that might hinder the efficiency of a conversation.

CCS CONCEPTS

• **Information systems → Personalization**.

KEYWORDS

Proactive Information Recall, Lifelogging, Conversational Lifelogs Retrieval

ACM Reference Format:
Yuan-Chi Lee, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2024. ConvLogRecaller: A Real-Time Conversational Lifelog Recaller. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3626772.3657659

1 INTRODUCTION

With the advance of instant messaging technology, people are used to chatting with family and friends in real-time over the Internet to exchange life experiences with each other. As the numerous events in our daily life, people cannot remember every detail. In other words, people may have difficulty in recalling the exact entities to text the messages, and even results in sending the wrong information. This explains the importance of an information recall system that helps people bring to mind what they are trying to recall. Suggesting the remaining words based on their life experiences to complete the sentence could be helpful for users to lower the burden of recalling their life events. Auto-completion is a common application as part of our everyday life. For instance, Gmail provides the service of interactively suggesting the responses in real-time to help users text messages quickly. While in everyday text, users may mention the food they ate last night or the names of friends who were at a party. Therefore, in addition to recommending standardized responses to help users complete their message, there is a need for an information recall system [5] that can detect the experiences users are recalling and suggest personalized responses drawn from their individual personal knowledge graph (PKG).

Generally, the information recall service can be offered in two modes: reactive and proactive [17]. In the reactive mode, users are required to actively request information for their recall needs. Jiang et al. [6] proposed a multimodal model that allows users to query their personal photo albums. A question-answering system was proposed by Yen et al. [18, 19] to either retrieve users' lifelogs or correct the unanswerable questions based on their personal knowledge bases. In addition, several studies have worked on visual lifelog retrieval [4, 12].

The proactive mode, which actively identifies the user's need for information recall, is rarely explored. Lin et al. [11] proposed the Structured Event Enhancement Network (SEEN) to detect the needs of information recall by comparing narratives with lifelogs. However, the lifelogs used in the work of Lin et al. [11] are documents, i.e., diaries. They only compare the differences between pairs of documents that describe the same individual's life experience to simulate the scenario of providing memory recall assistance. Different from previous works, we focus on a real-time conversational information recall system, ConvLogRecaller. [1] Specifically, in a conversation between a user and others, our system assists in timely recalling life experiences by offering a response for reference. When the user mixes up life events, the system immediately reminds the

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3657659

[1] https://energybubu.github.io/Recall/

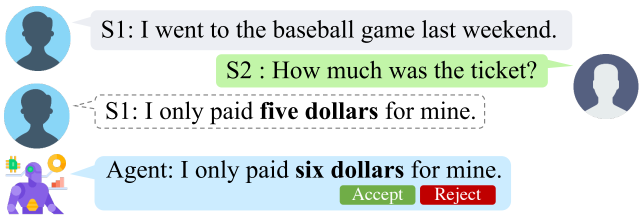


Figure 1: An Example of Information Recall Service

user and suggests a corrected description. For example, Figure 1 presents a daily conversation where the speaker (S1) discusses a life event with S2, but mistakenly recalls the ticket price. In this case, the agent (i.e., ConvLogRecall) detects the conflicts and offers a suggestion. S1 can determine whether to accept and send the suggested message, and can also directly ask for a suggested response. To sum up, the contributions in this paper are threefold:

(1) This work presents a new direction of lifelogging and demonstrates a personal intelligent system, ConvLogRecaller, which proactively provides the information recall service in the real-time conversation.
(2) We accomplish our goal with a large language model to suggest responses using retrieved lifelogs whenever users need to recall memory.
(3) We also construct a conversational lifelog dataset to detect users' information recall needs.[2]

2 RELEATED WORK

The issues of applying natural language processing techniques to assist users in text auto-completion have attracted much attention and have achieved great success. For example, the works of smart reply [7] and smart compose [2] provide contextual assistance to assist users to complete text when writing e-mails. However, these previous works only provide the service of recommending standardized responses without considering personal information. Yang et al. [16] attempt to recommend the next phrases to complete the sentence based on a PKG which consists of data about the things of mobile users' concerns, the people they contact, and the major activities they engaged in (e.g., meetings, exam, trip). The drawback of their work can be categorized into two parts. The first one is the relatively monotonous sources of lifelogs. They collect the clicking, adding, and deleting operations in apps, such as the contact list and calendars. The second one is limited text auto-completion services. In contrast to recommending possible next phrases to complete the messages only, the aim of this work is to explore advanced needs of information recall. This includes proactively aiding users in completing their messages and offering suggestions for correcting erroneous the wrong messages.

3 CONVLOGRECALLER

An overview of the design of the ConvLogRecaller is shown in Figure 2. The workflow in ConvLogRecaller is divided into two phases. The first one is the retrieval phase, which retrieves lifelogs related to the current conversation. The lifelogs are derived from

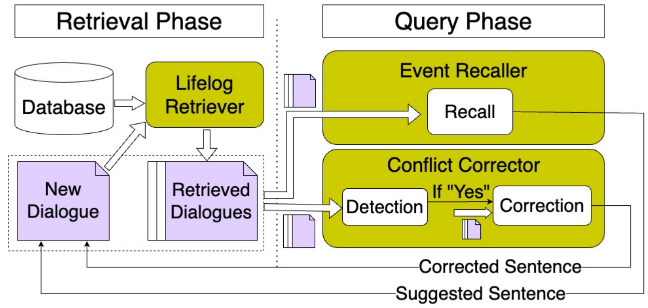


Figure 2: Overview of ConvLogRecaller

the user's past dialogues. Generally, information can come from any life logs rather than dialogues only. We build a lifelog retriever to retrieve the relevant dialogues. For the second phase, i.e., the query phase, the event recaller and conflict corrector provide the information recall and conflict detection services, respectively.

Lifelog Retriever: Given a current dialogue $\tilde{d} = \{t_1, t_2, ..., t_n\}$, where t_i is the i-turn in $\tilde{d}$ and n is the total number of turns, t_i is a speech from the j-th speaker $S_{j\in\{1,2\}}$. In this work, we set S_1 as the user. If S_1 requests recalling assistance to respond to t_n, our goal is to suggest t_{n+1}. Thus, the first step is to retrieve the relevant dialogues $\mathcal{D} = \{d_1, d_2, ..., d_m\}$ from S_1's dialogue database $\mathcal{B}$ based on $\tilde{d}$. We measure the cosine similarity between $\tilde{d}$ and each dialogue in $\mathcal{B}$ by ensembling two approaches, including term frequency-inverse document frequency (TF-IDF) and the dense passage retrieval (DPR) method proposed by Karpukhin et al. [8].[3] Formally, the similarity score is calculated by $\alpha \times F(Sim_{DPR}) + (1 - \alpha) \times F(Sim_{TF-IDF})$, where F denotes the normalization function that subtracts the mean and divides by the standard deviation for each element in the sentence representation, and α represents the weighted ratio. Then, we select top m dialogues as $\mathcal{D}$ based on similarity scores.[4]

Event Recaller: After retrieving $\mathcal{D}$, our system aims at producing t_{n+1} as requested by the user. Given $\mathcal{D}$, $\tilde{d}$, prompt $\mathcal{P}_r$, and a language model M, $t_{n+1} = M(\mathcal{D}, \tilde{d}; \mathcal{P}_r)$. The template of $\mathcal{P}_r$ is presented in Table 2.

Conflict Corrector: Apart from assisting users in recalling memory, our system also proactively detects the conflicts between the events mentioned in $\mathcal{D}$ and t_{n+1} written by the user S_1. If any conflicts are detected, our system will provide the corrections. We divide the correction process into two steps. The first step is conflict detection. We utilize M to detect whether a conflict exists. The detection result $y_{dc} = M(\mathcal{D}, \tilde{d}, t_{n+1}; \mathcal{P}_{dc})$, where $y_{dc} \in \{\text{Yes}, \text{No}\}$. If y_{dc} is "No", the system will not modify the t_{n+1}. Otherwise, we proceed to the second step, i.e., the conflict correction step. The system corrects the conflicts by providing the suggested $t_{n+1} = M(\mathcal{D}, \tilde{d}, t_{n+1}; \mathcal{P}_{sg})$. The templates of $\mathcal{P}_r$ and $\mathcal{P}_{sg}$ are shown in Table 2 and Table 3, respectively.

[2] https://github.com/ntunlplab/ConvLogRecaller-dataset

[3] The encoder in DPR is sentence transformer `all_minilm_l6_v2` [15]

[4] α and m are set to 0.8 and 10, respectively.

Table 1: Prompt template of event recaller

System Message: Based on the previous dialogues, please concisely provide the details.

User Message:
`[[ Previous Dialogues ]]`
1: {*Dialogues 1*}
2: {*Dialogues 2*}
3: {*Dialogues 3*}
.
.
.
`[[ New Dialogue ]]`
{*New Dialogue*}
`[[ Instruction ]]`
Finish what S1 should say, and provide details only based on previous dialogues in short.

Table 2: Prompt template of conflict detection

System Message: Detect conflicts between Previous Dialogues and New Dialogue, and then output "Yes" or "No".

User Message:
`[[ Previous Dialogues ]]`
1: {*Dialogues 1*}
2: {*Dialogues 2*}
3: {*Dialogues 3*}
.
.
.
`[[ New Dialogue ]]`
{*New Dialogue*}
`[[ Instruction ]]`
Does New dialogue conflict with Previous dialogues?

4 DATASET CONSTRUCTION

We annotate several dialogues $\tilde{D} = \{\tilde{d}_1, \tilde{d}_2, ..., \tilde{d}_r\}$ to simulate the scenario of information recall. Specifically, the k-th $\tilde{d}_k$ is annotated according to $\mathcal{D}_k$. Hence, we construct $\mathcal{D}_k$ first by manually selecting some dialogues from the LED dataset [3]. In LED, 1,003 dialogues were sampled from DailyDialog [9] and annotated with speakers' life events. The selected $\mathcal{D}_k$ is viewed as the k-th S_j's dialogue history. However, selecting dialogues to construct $\mathcal{D}_k$ from LED is challenging due to mismatches in speaker's life events. Thus, we use GPT-3.5 [14] to generate 5 dialogues as $\mathcal{D}_k$ by a given topic and an event. We also use Falcon [1] to paraphrase $\mathcal{D}_k$ to increase lexical diversity. For the t^k_{n+1} turn in $\tilde{d}_k$, S_1 may request the recall service, present a statement conflicting with $\mathcal{D}_k$, or exhibit no particular need in our design. Finally, we construct 211 dialogues ($r = 211$),[5] where 71 and 140 dialogues are used for the evaluations of the event recaller and the conflict corrector, respectively.

[5]We extend the use of publicly available dialogue datasets, which do not contain any personal information about the speakers, thereby ensuring there are no ethical issues.

Table 3: Prompt template of conflict correction

System Message: Precisely determine conflicts between `[[ New Dialogue ]]` and `[[ Previous Dialogues ]]`.

User Message:
`[[ Previous Dialogues ]]`
1: {*Dialogues 1*}
2: {*Dialogues 2*}
3: {*Dialogues 3*}
.
.
.
`[[ New Dialogue ]]`
{*New Dialogue*}
`[[ Instruction ]]`
Concisely point out the conflict between New Dialogue and Previous Dialogues precisely.

Assistant Message:
{*The answer from the assistant*}

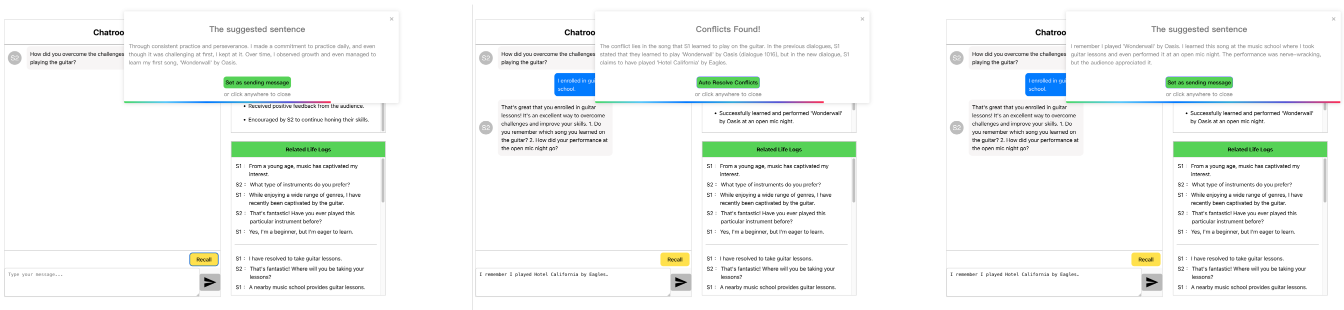
User Message:
Rewrite the last utterance in New Dialogue only based on Previous Dialogues to resolve Conflicts and provide details as more as you can.

Table 4: Experimental results

Model	Event Recaller	Conflict Detection	Conflict Correction
GPT-3.5	0.2262	34.29%	40.00%
GPT-4	0.1808	89.29%	82.86%

5 EXPERIMENTS

We adopt either the `gpt-3.5-turbo` or the `gpt-4-0613` [13] models as the backbone language models in experiments. The results are shown in Table 4. We evaluate the performances of the event recaller with the ROUGE-1 F_1 [10]. The conflict detection and correction are manually evaluated by human annotators. Three annotators are invited to evaluate the conflict detection and correction results. The criteria are whether the model detects conflicts between the generated t_{n+1} and $\mathcal{D}$, and whether it accurately corrects these conflicts, respectively. The inter-annotator agreement for the correction results of `gpt-3.5-turbo` and `gpt-4-0613`, calculated using Fleiss' Kappa, are 0.7619 and 0.7166, respectively. `gpt-4-0613` seems to be less effective as the event recaller, this is primarily due to differences in wording between the generated t_{n+1} and the annotated sentence. In the conflict correction, we found that `gpt-3.5-turbo` struggles to detect the conflicts and it usually incorporates the conflicts into the suggested correction. For instance, if a user mistakenly recalls wearing size 32 shoes as size 30, the model suggests the correction as "I used to wear shoes in size 32, but now in 30", rather than correcting the size to 30. Overall, `gpt-4-0613` achieves promising results.



(a) An Example of Event Recalling **(b) An Example of Conflict Detection** **(c) An Example of Conflict Resolution**

Figure 3: Screenshot of ConvLogRecall's Online Service Interface

6 DEMONSTRATION

Figure 3 shows the screenshot of ConvLogRecaller's service interface.[6] We present a chat platform to simulate conversational scenarios. Within the Conversation block, users can input messages as the speaker named "S1". We also provide users with lifelogs of several speakers, enabling them to role-play as a specific speaker (referred to as S1 in conversations) and mimic daily conversations with others, talking about past life experiences.

In the context shown in Figure 3a, the conversation initiates when S2 poses a question to S1 about how she/he learned to play the guitar and overcame challenges. The experiences of S1 learning the guitar involve taking lessons at a nearby music school and successfully mastering "Wonderwall by Oasis" to perform at an open mic night. If S1 wants to respond to S2 but finds her/himself struggling to recall specific details of the event, S1 can press the yellow button "Recall". Our system then generates suggested sentences based on the question and S1's lifelogs, displaying the sentences in a pop-up window in the top-right corner of the system interface, offering these as references for S1 (i.e., the user). After verifying the accuracy of the suggested sentences, users can decide whether to send the suggested sentences as their response or to enter their messages.

Our system also automatically detects inconsistencies between a user's narrative and their lifelogs. If discrepancies are found, it notifies the user of the inconsistency and generates corrected sentences. As shown in Figure 3b, after successfully responding to S2's initial question, S2 further inquires about S1's experience performing with the guitar. S1 incorrectly claims to have played "Hotel California by the Eagles" instead of the correct performance of "Wonderwall by Oasis at an open mic night". Upon detecting the inconsistency, our system will inform the user, highlight the conflict, and provide the relevant lifelogs to review. This assists users in recalling the relevant experience effectively. Subsequently, users have the option to allow the system to assist in correcting descriptions of their life experiences. In Figure 3c, by clicking the "auto resolve conflicts" button, our system generates the corrected sentences, enabling users to directly send this message.

[6]The video of our system demonstration is available at: https://drive.google.com/file/d/15puaP0z9Ere8ctG8lzG4unj84rf2E_-a/view?usp=sharing

7 CONCLUSION AND FUTURE WORK

This work demonstrates a pilot conversational information recall system, ConvLogRecaller, to help people recall their memories during real-time daily conversations. Due to the scarcity of publicly available datasets that record personal exchanges, we constructed a dataset that mimics individuals sharing their past life experiences. Moreover, our dataset also consists of conversations that simulate situations where participants occasionally forget or misremember events. Although the size of our dataset is limited, it serves as a foundational step for investigating the mechanisms of information recall support in this pilot study. Compared with other lifelog retrieval systems, our system can not only support users to recall their experiences, but also proactively suggest the corrected messages if the target user mixes up the life events. We plan to develop an advanced retrieval-augmented language model into our system to improve the performances of event recalling and conflict detection services. In addition, constructing a privacy-aware framework that prevents disclosing users' data to the global client is left as the future work.

8 LIMITATIONS

This work explores the application of an information recall assistant within a chat application context. However, the current study still exhibits several limitations and unresolved issues that warrant further investigation. For instance, the management of personal knowledge graphs or lifelog data, which should ideally be decentralized to enhance user privacy, has not yet been addressed in detail. Another significant challenge is the occurrence of hallucinations within large language models, which can undermine the accuracy of recalled information and potentially erode user trust. Although we currently do not have a solution to completely eliminate hallucinations, our system provides suggestions that users can review and decide whether to accept, aiming to mitigate some of the potential issues arising from this problem. These limitations highlight areas for further investigation.

ACKNOWLEDGMENTS

This research was partially supported by National Science and Technology Council, Taiwan, under grants MOST 110-2221-E-002-128-MY3 and NSTC 112-2634-F-002-005-.

REFERENCES

[1] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance. (2023).

[2] Mia Xu Chen, Benjamin N Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M Dai, Zhifeng Chen, et al. 2019. Gmail smart compose: Real-time assisted writing. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2287–2295.

[3] Yi-Pei Chen, An-Zi Yen, Hen-Hsen Huang, Hideki Nakayama, and Hsin-Hsi Chen. 2023. LED: A Dataset for Life Event Extraction from Dialogs. *arXiv preprint arXiv:2304.08327* (2023).

[4] Tzu-Hsuan Chu, Hen-Hsen Huang, and Hsin-Hsi Chen. 2019. Image recall on image-text intertwined lifelogs. In *IEEE/WIC/ACM International Conference on Web Intelligence*. 398–402.

[5] Cathal Gurrin, Björn Þór Jónsson, Duc Tien Dang Nguyen, Graham Healy, Jakub Lokoc, Liting Zhou, Luca Rossetto, Minh-Triet Tran, Wolfgang Hürst, Werner Bailer, et al. 2023. Introduction to the Sixth Annual Lifelog Search Challenge, LSC'23. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*. 678–679.

[6] Lu Jiang, Junwei Liang, Liangliang Cao, Yannis Kalantidis, Sachin Farfade, and Alexander Hauptmann. 2017. Memexqa: Visual memex question answering. *arXiv preprint arXiv:1708.01336* (2017).

[7] Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, Laszlo Lukacs, Marina Ganea, Peter Young, et al. 2016. Smart reply: Automated response suggestion for email. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 955–964.

[8] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 6769–6781.

[9] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Asian Federation of Natural Language Processing, Taipei, Taiwan, 986–995. https://aclanthology.org/I17-1099

[10] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[11] You-En Lin, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2022. SEEN: Structured Event Enhancement Network for Explainable Need Detection of Information Recall Assistance. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 5438–5451.

[12] Thao-Nhu Nguyen, Tu-Khiem Le, Van-Tu Ninh, Cathal Gurrin, Minh-Triet Tran, Thanh Binh Nguyen, Graham Healy, Annalina Caputo, and Sinead Smyth. 2023. E-LifeSeeker: An Interactive Lifelog Search Engine for LSC'23. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge*. 13–17.

[13] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[14] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[15] 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*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[16] Yu Yang, Jiangxu Lin, Xiaolian Zhang, and Meng Wang. 2022. PKG: A personal knowledge graph for recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3334–3338.

[17] An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. Ten questions in lifelog mining and information recall. In *Proceedings of the 2021 international conference on multimedia retrieval*. 511–518.

[18] An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. Unanswerable question correction in question answering over personal knowledge base. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14266–14275.

[19] An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2022. Unanswerable Question Correction and Explanation over Personal Knowledge Base. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 4645–4649.