Overview of NTCIR-18 Transfer-2 Task

Hideo Joho[1], Atsushi Keyaki[2], Yuuki Tachioka[3], Shuhei Yamamoto[1]

[1]University of Tsukuba, Japan [2]Hitotsubashi University, Japan [3]Denso IT Laboratory, Japan

Introduction

The first round of the NTCIR Transfer Task, called Transfer-1, was a pilot task held at NTCIR-17. The outcomes of Transfer-1 demonstrated the potential of dense retrieval methods for Japanese, but also underscored the need for more diverse and comprehensive datasets to address the full range of challenges in the domain. Based on the insight from the first round, Transfer-2 at NTCIR-18 sets two subtasks: Retrieval Augmented Generation (RAG) and Dense Multimodal Retrieval (DMR). A total of 14 runs was submitted from two teams and organiser.

Retrieval Augmented Generation (RAG)

RAG subtask aims to develop a retrieval module suitable for Retrieval-Augmented Generation (RAG). RAG utilizes external knowledge retrieved by a retrieval module during response generation by an LLM to produce high-quality responses. At Transfer-2, we focused on the open-domain factoid question answering (QA) task for the RAG subtask because facts are presented in response to questions in the open-domain factoid QA task, making evaluation easier.

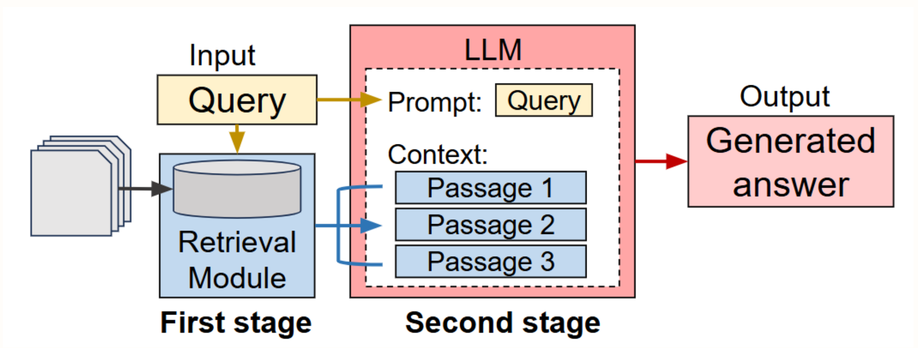

Figure 1: Overview of the RAG subtask.

We used the AIO Official Dataset Version 2.0 which contains passage-level relevance assessments for the retriever and answer sets for the reader in the question-answering stage, which aligns with the goals of our subtask. The dataset includes 22,335 QA pairs for training, 1,000 QA pairs for development, and 1,000 questions for testing. The target corpus for RAG consists of Wikipedia articles

Additionally, we have adopted a two-stage retrieval model for the open-domain factoid QA task, specifically the retriever-reader model used in open-domain QA systems. For the first stage, the input is a natural language question, and the output is the top 100 ranked passage IDs corresponding to the natural language question. We use training data and development and evaluation data from the AIO Official Dataset Version 2.0 for training and evaluation, respectively. The evaluation metrics are meanAveragePrecision(mAP), HitRate(HR)@k ($\mathrm{k} = 1, 5, 10, 50, 100$) and nDCG@k ($\mathrm{k} = 1, 5, 10, 50, 100$).

For the second stage, the input is a natural language question and the k passages retrieved in the first stage, and the output is the answer to the natural language question. The evaluation metric is the accuracy. The baseline model provided by the organizers adopts the Fusion-in-Decoder (FiD), a model specialized for open-domain QA tasks.

Dense Multimodal Retrieval (DMR)

The DMR subtask is deinfed as follows: given a non-linguistic modality query $\mathbf{q}$ and a corpus $\{\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_\mathrm{n}\}$ consisiting of n data points in another modality, the goal of cross-modal retrieval is to find the k data points that are most relevant to the query $\mathbf{q}$, where $\mathrm{k} \ll \mathrm{n}$. The task is to efficiently generate the top-k candidates that are relevant to the query based on the similarity metric $\mathrm{sim}(\mathbf{q}, \mathbf{d}) \in \mathbb{R}$. In this context, the query $\mathbf{q}$ and data $\mathbf{d}$ can refer to sensor data $\{\mathbf{q}^{\mathrm{sen}}, \mathbf{d}^{\mathrm{sen}}\} \in \mathbb{R}^S$ or image data $\{\mathbf{q}^{\mathrm{img}}, \mathbf{d}^{\mathrm{img}}\} \in \mathbb{R}^{3 \times W \times \mathrm{H}}$, where S denotes the number of dimensions of the sensor data and W and H represent the width and height of the image data. The evaluation metric is the mean reciprocal rank (MRR).

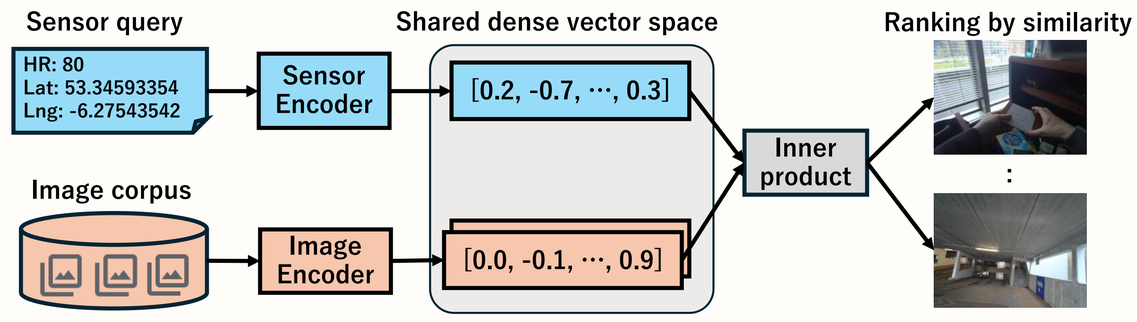

Figure 2: Overview of the DMR subtask.

To effectively advance this subtask, we reuse Lifelog Search Challenge 2024 (LSC'24) dataset [**?**], one of the largest multi-modal datasets derived from users' daily activities. The dataset was generated by one active lifelogger and is 18 months in length. It includes non-linguistic modalities such as ego-centric images, heart rate, and location information (e.g., latitude/longitude), providing a robust foundation for testing our dense retrieval approach in a real-world, multi-modal context. Thereby, deemed conducive to the realization of the subtask's objectives.

We created 140 retrieval topics for the formal run. The topics were extracted from the period from March 1, 2020, to June 30, 2020. Each topic is generated on a daily basis, and sensors or images recorded at the same time as the query are extracted as relevant (correct) data. Thus, the retrieval target for each topic is the set of data recorded on each day, with only one relevant data set and 199 irrelevant data sets for each topic. Additionally, we designed two types of retrieval tasks: one that retrieves images from sensors (reffered to as *sen2img*) and another that retrieves sensors from images (reffered to as *img2sen*). Examples of created topics are shown in Figure 3. Note that the sensor data were normalized in all topics.

Topic ID	Timestamp	Query	Relevant Data
3*img2sen_0309	3*2020-03-09 08:39:24	3*20200309_083924.jpg	hr: 1.891 lat: -0.121 lng: 0.055
3*sen2img_0626	3*2020-06-26 20:47:05	hr: 0.015 lat: -0.121 lng: 0.055	3*20200626_204705.jpg

20200309_083924.jpg

20200626_204705.jpg

Figure 3: Examples of created topics. Timestamps denote the time when the data was recorded. "hr", "lat", and "lng" represent heart rate, latitude, and longitude, respectively.

The organizing team developed two baseline runs and conducted evaluations with 10 and 100 training iterations (ORG:baseline-10 and ORG:baseline-100).

Participated Systems and Results

We have four runs for RAG subtask and ten runs for DMR subtask. Please visit individual poster presentations for the details of participated systems and their results!

Conclusions & Acknowledgements

In conclusion, the Transfer-2 task provided valuable benchmarks, resources, and insights for future research on retrieval augmented applications and cross-modal information access. We thank the orgnizers of AIO and NTCIR-18 Lifelog-6 Task for making their datasets available to us.