NTCIR-18: Proceedings of the 18th NTCIR Conference on Evaluation ofInformation Access Technologies, June 10-13, 2025, Tokyo, Japan
DOI: https://doi.org/10.20736/0002002081

Overview of the NTCIR-18 Transfer-2 Task

Hideo Joho
University of Tsukuba
hideo@slis.tsukuba.ac.jp

Atsushi Keyaki
Hitotsubashi University
a.keyaki@r.hit-u.ac.jp

Yuuki Tachioka
Denso IT Laboratory
tachioka.yuki@core.d-itlab.co.jp

Shuhei Yamamoto
University of Tsukuba
syamamoto@slis.tsukuba.ac.jp

Abstract

This paper provides an overview of the NTCIR-18 Transfer-2 task that aims to bring together researchers from Information Retrieval, Machine Learning, and Natural Language Processing to develop a suite of technology for transferring resources generated for one purpose to another in the context of dense retrieval. Two subtasks were run for this round: the Retrieval Augmented Generation (RAG) subtask and the Dense Multimodal Retrieval (DMR) subtask. This paper presents the dataset developed and evaluation results of participant runs. Note that this paper includes material from our earlier work published in [3], revised for the current work.

Keywords

Retrieval Augmented Generation, Cross-Modal Retrieval, Test Collection

Subtasks

Retrieval Augmented Generation (Japanese)
Dense Multimodal Retrieval (English)

1 Introduction

The first round of the NTCIR Transfer Task, called Transfer-1 [1], was a pilot task held at NTCIR-17 [2]. It consisted of two main subtasks: Dense First-Stage Retrieval and Dense Reranking. The aim of Transfer-1 was to evaluate dense retrieval methods specifically for Japanese information retrieval, addressing challenges related to orthographic diversity and retrieval effectiveness. 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[1] at NTCIR-18 [4] sets two subtasks: Retrieval Augmented Generation (RAG) and Dense Multimodal Retrieval (DMR).

2 Task Design

2.1 RAG subtask

This 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. According to the study that analyzed the performance impact of the retrieval module [15], there was a difference of more than 30% in performance depending on the selection of documents used for RAG. Interestingly, RAG behaves differently from typical retrieval systems. That is, related documents (documents related to the query but not containing the correct answer) decrease the performance of RAG and have a worse impact than irrelevant documents. From this, it is clear that the strategy for retrieving documents within the RAG framework is a significant factor. In this subtask, we will explore a retrieval module suitable for RAG together with subtask participants. The ideal design of RAG prompts, such as the number of documents given to the generator, is also a focus.

There are two main issues to consider when carrying out this subtask: 1. The design of input/output and evaluation, and 2. The design of a feasible development environment for the participants. With respect to 1, fixing the query and searching the target documents is necessary to ensure a fair comparison. In addition, careful consideration is needed to find a method for evaluating LLMs that allows reproducibility. With respect to 2, it is necessary to use a practical and high-performance LLM to generate responses that withstand verification. However, the problem is that operating such a high-performance LLM requires high-performance GPUs.

Based on these considerations, we have decided to focus 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. 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. In the first stage, the retriever retrieves documents or passages that are candidates for the answer's evidence from a large corpus, similar to a standard information retrieval task. For the retriever, we expect to use classical sparse vector search models such as BM25 or popular dense vector search models such as Dense Passage Retrieval (DPR) [9]. Another option is to use more advanced dense vector search models that participants used in Transfer-1.

In the second stage, the reader extracts an answer from the set of relevant documents or passages retrieved by the retriever. The baseline model provided by the organizers adopts the Fusion-in-Decoder (FiD)[2] [16], a model specialized for open-domain QA tasks. Note that FiD is not as large as cutting-edge LLMs, which makes it more accessible to many participants.

Another baseline provided by the organizers is the LlamaIndex one. LlamaIndex is the widely used general-purpose RAG pipeline. LlamaIndex also adopts the retriever-reader model, which is essentially the same as the above baseline, but the use of LlamaIndex makes RAG tasks easier for nonexperts.

We believe that we can achieve our objectives by evaluating the performance and analyzing the trends of both the retriever

[1]Homepage https://github.com/ntcirtransfer/transfer2/discussions

[2]https://github.com/facebookresearch/FiD

and the reader. We used the AIO Official Dataset Version 2.0[3] to reduce corpus construction costs. The AIO Official Dataset Version 2.0 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[4].

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 ($k = 1, 5, 10, 50, 100$) and nDCG@k ($k = 1, 5, 10, 50, 100$). 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.

2.2 DMR subtask

2.2.1 Overview. This subtask aims to encourage participants to develop technologies for the retrieval of suitable data across diverse modalities such as images, sensor information, and location data. Specifically, the task involves the transfer from a source modality to a target modality. Participants are tasked with engineering technologies capable of representing multiple modalities within a shared dense vector space.

One key challenge in implementing this subtask is how to determine positive and hard negative samples for contrastive learning [10], particularly due to the lack of established datasets paired with non-linguistic modalities. Despite the advancements in LLMs that have paved the way for cross-modal information processing, enabling interactions across various modalities, including text, images, and audio, the methodology required for effective cross-modal information access remains unclear [11]. Although large text and image corpora exist (e.g., Microsoft COCO [12] and Flickr30k [13]), they often lack the corresponding data from other modalities, limiting their utility in developing robust cross-modal information access technologies. Additionally, it remains a challenge to design models capable of processing non-linguistic modalities, such as sensor data (e.g., heart rate and accelerometer ratings) or location information (e.g., latitude and longitude), which are crucial for considering user context.

2.2.2 Task Definition. An Overview of the DMR subtask is shown in Figure 1 [19]. The DMR subtask is deinfed as follows: given a non-linguistic modality query $\mathbf{q}$ and a corpus $\{\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_n\}$ 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 $k \ll n$. The task is to efficiently generate the top-k candidates that are relevant to the query based on the similarity metric $\text{sim}(\mathbf{q}, \mathbf{d}) \in \mathbb{R}$. In this context, the query $\mathbf{q}$ and data $\mathbf{d}$ can refer to sensor data $\{\mathbf{q}^{\text{sen}}, \mathbf{d}^{\text{sen}}\} \in \mathbb{R}^S$ or image data $\{\mathbf{q}^{\text{img}}, \mathbf{d}^{\text{img}}\} \in \mathbb{R}^{3\times W\times H}$, 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).

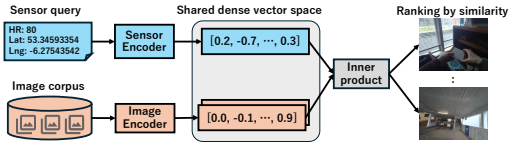


Figure 1: Overview of the DMR subtask.

2.2.3 Dataset and Topics. To effectively advance this subtask, we reuse Lifelog Search Challenge 2024 (LSC'24) dataset [14], 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.

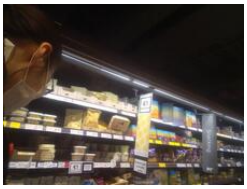
We created 140 retrieval topics for the formal run. The topics were extracted from the period from March 1, 2020, to June 30, 2020. Each topic is generated on a daily basis, and sensors or images recorded at the same time as the query are extracted as relevant (correct) data. Thus, the retrieval target 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 2. Note that the sensor data were normalized in all topics.

2.2.4 Baseline. The baseline system was provided to the participants by the organizers[5]. To embed image and sensor data into a shared dense vector space, we build two types of deep neural network (DNN) encoders f and g. Specifically, for the image encoder f, we utiliize two ResNet18 [20] models pretrained on the ImageNet [21] and Places365 [22] datasets. The feature vectors obtained through the pretrained models were combined and transformed into a 512-dimensional feature vector using a single multilayer perception (MLP) layer. The numerical sensor data, such as heart rate, were normalized so that each data distribution had a mean of 0 and a standard deviation of 1. Categorical sensor data, such as semantic location information, were transformed using frequency encoding. Subsequently, these data were transformed into 512-dimensional vectors through two MLP layers as g. Using these encoders, we obtain two dense vectors: $g(\mathbf{q}^{\text{sen}}) = \mathbf{h}_q^{\text{sen}}$ and $f(\mathbf{d}^{\text{img}} = \mathbf{h}_d^{\text{img}})$. The relevance of $\mathbf{q}^{\text{sen}}$ to $\mathbf{d}^{\text{img}}$ is computed as the dot product of their corresponding dense representations as $\text{sim}(\mathbf{q}^{\text{sen}}, \mathbf{d}^{\text{img}}) = \langle \mathbf{h}_q^{\text{sen}}, \mathbf{h}_d^{\text{img}} \rangle$. Similarly, the relevance of $\mathbf{q}^{\text{img}}$ to $\mathbf{d}^{\text{sen}}$ is computed using the same process.

[3] https://sites.google.com/view/project-aio/dataset
[4] https://github.com/cl-tohoku/quiz-datasets
[5] https://github.com/ntcirtransfer/transfer2/tree/main/DMR

Topic ID	Timestamp	Query	Relevant Data
img2sen_0309	2020-03-19 09:03:00	20200309_083924.jpg	hr: 1.891 lat: -0.121 lng: 0.055
sen2img_0626	2020-06-26 20:47:00	hr: 0.015 lat: -0.121 lng: 0.055	20200626_204705.jpg

20200309_083924.jpg



20200626_204705.jpg

Figure 2: Examples of created topics. Timestamps denote the time when the data was recorded. "hr", "lat", and "lng" represent heart rate, latitude, and longitude, respectively.

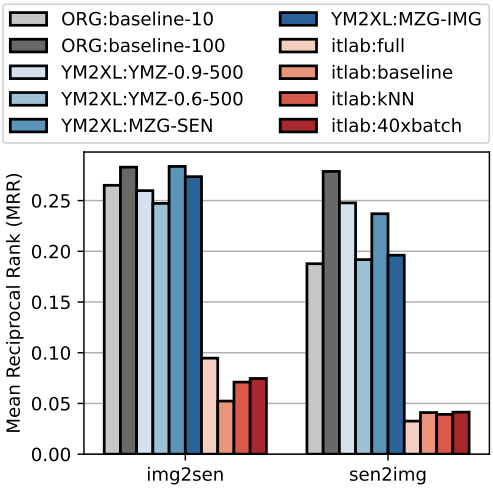


Figure 3: Mean reciprocal rank (MRR) of each run.

3 Participated Systems

3.1 RAG subtask

Table 1 shows the evaluation metrics for the first stage and the second stage. ditlab team participated in this subtask [18]. First, they developed two baseline systems: the first is the LlamaIndex baseline (run1) and the second is the DPR and Llama baseline (run2). Second, based on the DPR, they finetune LLM by low-rank approximation (run3), and they developed a late fusion of multiple answers generated by RAG that use multiple contexts (run4). They found that performance of the second stage was similar for both baselines and although the retrieved documents are the same, finetuning improved performance and late fusion was effective for this QA task.

3.2 DMR subtask

3.2.1 YMX2L. The team of YMX2L [23] submitted four runs for the DMR subtask, employing image/sensor data augmentation (YMX2L:MZG-IMG and YMX2L:MZG-SEN) and object detection (YMX2L:YMZ-0.9-500 and YMX2L:YMG-0.6-500). Image data augmentation is achieved by a combination of geometric transformations, and sensor data augmentation is achieved by adding noise (Jittering) and constant doubling (Scaling). The results of object detection using YOLOv11 [24] are represented by a bag of objects, which is converted into a feature vector by means of an MLP layer. These techniques are realized by adding them to the organizer baseline.

3.2.2 ditlab. The ditlab team [18] submitted four runs for the DMR subtask, developing a modality-aware sensor encoder and geolocation encoder. For categorical features in the sensor data, their sensor encoder first converted them into prompt-style text strings and then encoded them using Sentence-BERT. The numerical and textual embeddings are concatenated and passed through another MLP to project the sensor record into the same space as the image embeddings. The geolocation encoder converts raw coordinates (i.e., latitude and longitude) into textual address strings to capture semantic cues.

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

3.2.4 Evaluation results. Figure 3 shows the MRR of each run. In the *img2sen* task, the YMX2L:MZG-SEN run achieved the best performance, while in the *sen2img* task, the ORG:baseline-100 run performed the best. The breakdown of reciprocal rank by topic is shown in Figure 4. The ORG and YMX2L teams' runs show similar reciprocal ranks on most topics. It is suggested that this is because the YMX2L team's submission system was developed based on the organizer's baseline. Interestingly, the ditlab team's runs are sometimes successfully estimated on topics where the ORG and YMX2L team's runs are not properly estimated (e.g., *img2sen_0503* and *sen2img_0509*). This is suggested to be a case where the two types of encoders developed by ditlab (modality-aware encoder and geolocation encoder) are functioning properly. By analyzing the characteristics of topics that differ in accuracy, we believe that we can gain insight into the development of more accurate cross-modal search technology.

4 Conclusions

This paper presented an overview of the NTCIR-18 Transfer-2 task, which focused on advancing research on resource transfer within the context of dense retrieval through two subtasks: Retrieval Augmented Generation (RAG) and Dense Multimodal Retrieval (DMR) subtasks. For the RAG subtask, we introduced an open-domain factoid QA framework using a two-stage retriever-reader model. The evaluation of the participant systems revealed the effectiveness of fine-tuning and late-fusion strategies in improving the accuracy of the answers, highlighting the importance of quality of retrieval in RAG pipelines. For the DMR subtask, we prepared the challenge of cross-modal retrieval involving non-linguistic data such as sensor and image modalities, leveraging the LSC'24 lifelog

Table 1: Evaluation metrics for the first stage (mAP, NDCG, and HR) and those of the second stage (Acc) of the RAG subtask, where run1 is the LlamaIndex baseline, run2 is the DPR and Llama baseline, run3 is the LLM with LoRA finetuning for open-domain QA, and run4 is the late fusion of top-7 outputs by RAG with the seven contexts with the seven highest similarity scores.

	mAP	NDCG@1	NDCG@5	NDCG@10	NDCG@50	NDCG@100	HR@1	HR@5	HR@10	HR@50	HR@100	Acc
ditlab												
dev set												
run1	0.045	0.241	-	-	-	-	0.241	-	-	-	-	0.300
run2	0.185	0.388	0.304	0.289	0.318	0.342	0.388	0.625	0.711	0.844	0.873	0.297
run3	0.185	0.388	0.304	0.289	0.318	0.342	0.388	0.625	0.711	0.844	0.873	0.357
run4	0.185	0.388	0.304	0.289	0.318	0.342	0.388	0.625	0.711	0.844	0.873	0.404
test set												
run1	0.407	0.407	-	-	-	-	0.407	-	-	-	-	0.391
run2	0.319	0.461	0.389	0.375	0.462	0.565	0.461	0.684	0.767	0.878	0.918	0.395
run3	0.319	0.461	0.389	0.375	0.462	0.565	0.461	0.684	0.767	0.878	0.918	0.437
run4	0.319	0.461	0.389	0.375	0.462	0.565	0.461	0.684	0.767	0.878	0.918	0.503

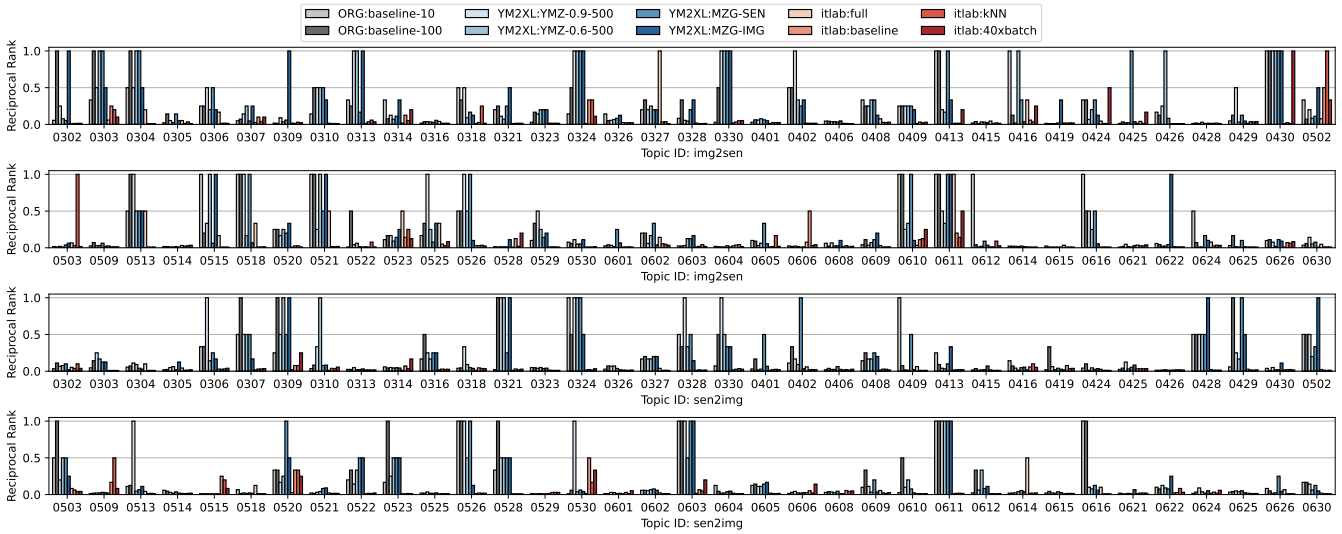


Figure 4: Reciprocal rank of each run by topic.

dataset. Among the participating teams, YMX2L employed data augmentation techniques and object detection to improve retrieval performance, achieving the best MRR in the *img2sen* task. The ditlab team proposed modality-aware and geolocation-aware encoders, showing strength in specific topics where other systems struggled. In conclusion, the Transfer-2 task provided valuable benchmarks, resources, and insights for future research on retrieval augmented applications and cross-modal information access.

Acknowledgements

We thank participants of Transfer-2 task for their interest and dedication to develop novel technologies. We also thank the orgnizers of NTCIR-18 Lifelog-6 Task for making their valuable datasets available to us. Finally, we thank the NTCIR office and orgaizers of NTCIR-18 for their continuous support.

References

[1] Hideo Joho, Atsushi Keyaki, and Yuki Ohba. (2023). Overview of the NTCIR-17 Transfer Task. In Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, December 12-15, 2023, Tokyo, Japan. https://doi.org/10.20736/0002001319

[2] Takehiro Yamamoto and Zhicheng Dou. (2023). Overview of the NTCIR-17. In Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, December 12-15, 2023, Tokyo, Japan. https://doi.org/10.20736/0002001332

[3] Hideo Joho, Atsushi Keyaki, Yuuki Tachioka and Shuhei Yamamoto. (2024). Building Test Collections for Japanese Dense Information Retrieval Technologies and Beyond. In Joint Proceedings of the 1st Workshop on Evaluation Methodologies, Testbeds and Community for Information Access Research (EMTCIR 2024) and the 1st Workshop on User Modelling in Conversational Information Retrieval (UM-CIR 2024), co-located with the 2nd International ACM SIGIR Conference on Information Retrieval in the Asia Pacific (SIGIR-AP 2024).

[4] Qingyao Ai, Chung-Chi Chen, and Shoko Wakamiya. (2025). Overview of the NTCIR-18. In Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies, June 10-13, 2025, Tokyo, Japan.

[5] Craig Macdonald and Nicola Tonellotto. (2020). Declarative Experimentation in Information Retrieval using PyTerrier. In Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval (ICTIR '20).

Association for Computing Machinery, New York, NY, USA, 161–168. https://doi.org/10.1145/3409256.3409829

[6] Kazuma Takaoka and Sorami Hisamoto and Noriko Kawahara and Miho Sakamoto and Yoshitaka Uchida and Yuji Matsumoto. (2018) Sudachi: a Japanese Tokenizer for Business. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). https://aclanthology.org/L18-1355

[7] Omar Khattab and Matei Zaharia. (2020). ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20). Association for Computing Machinery, New York, NY, USA, 39–48. https://doi.org/10.1145/3397271.3401075

[8] Luiz Henrique Bonifacio, Israel Campiotti, Roberto de Alencar Lotufo, and Rodrigo Frassetto Nogueira. (2021). mMARCO: A Multilingual Version of MS MARCO Passage Ranking Dataset. arXiv:2108.13897

[9] 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), pages 6769–6781, Online. Association for Computational Linguistics.

[10] Wu, Z., Xiong, Y., Yu, S.X., Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3733–3742.

[11] Wang, K., Yin, Q., Wang, W., Wu, S., Wang, L. (2016). A comprehensive survey on cross-modal retrieval. arXiv preprint arXiv:1607.06215.

[12] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L. (2014). Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755.

[13] Young, P., Lai, A., Hodosh, M., Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics **2**, 67–78 (2014)

[14] Gurrin, C., Zhou, L., Healy, G., Bailer, W., Dang Nguyen, D.T., Hodges, S., Jónsson, B.T., Lokoč, J., Rossetto, L., Tran, M.T., Schöffmann, K. (2024). Introduction to the seventh annual lifelog search challenge, lsc'24. In: Proceedings of the 2024 International Conference on Multimedia Retrieval. p. 1334–1335. ICMR'24, Association for Computing Machinery, New York, NY, USA.

[15] Cuconasu, F., Trappolini G., Siciliano, F., Filice, S. (2024). The Power of Noise: Redefining Retrieval for RAG Systems, arXiv:2401.14887.

[16] Izacard, G., Grave, E. (2020). Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering, arXiv:arXiv:2007.00128.

[17] Shijie Wu and Mark Dredze. (2020). Are All Languages Created Equal in Multilingual BERT?. In Proceedings of the 5th Workshop on Representation Learning for NLP, pages 120–130, Online. Association for Computational Linguistics.

[18] Tachioka, Y and Terao, Y. (2025). ditlab at the NTCIR-18 Transfer-2 Task. In Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies, June 10-13, 2025, TOkyo, Japan.

[19] Yamamoto, S and Kando, N. (2025). Temporal Closeness for Enhanced Cross-Modal Retrieval of Sensor and Image Data, In proceedings of the MMM2025, pp. 170–183, Nara, Japan.

[20] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In proceedings of the CVPR'16, pp. 770–778.

[21] Krizhevsky, A., Sutskever, I., Hinton, G.E. (2012). ImageNet classification with deep con- volutional neural networks. In proceedings of the NIPS'12. vol. 25.

[22] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places (2018). A 10 million database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(6), 1452–1464.

[23] Mizuguchi, R., Yamazaki, T., and Yamamoto, S. (2025). YMX2L at the NTCIR-18 Transfer-2 Task. In Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies, June 10-13, 2025, TOkyo, Japan.

[24] Khanam, R. and Hussain, M. (2024). Yolov11: An overview of the key architectural enhancements. arXiv preprint arXiv:2410.17725.