Building Test Collections for Japanese Dense Information Retrieval Technologies and Beyond

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

[1]Institute of Library, Information and Media Science, University of Tsukuba
[2]Graduate School of Social Data Science, Hitotsubashi University
[3]Denso IT Laboratory

Abstract

This paper presents the NTCIR Transfer Task, a series of evaluations aimed at advancing dense information retrieval technologies for non-English languages, particularly Japanese. While dense retrieval methods have shown significant advancements for English-language content, their application to Japanese remains limited due to a lack of dedicated datasets and resources. The Transfer Task aims to address this gap by building specialized test collections and organizing focused tasks, such as Retrieval-Augmented Generation, Dense Multimodal Retrieval, and Dense Cross-Language Retrieval. The first round, Transfer-1, focused on dense retrieval and reranking for Japanese, while Transfer-2 expands to incorporate cross-lingual and multimodal tasks. By fostering the development of effective ranking technologies for Japanese and cross-lingual contexts, this research contributes to enhancing global information access and promotes equitable information retrieval capabilities for diverse languages and modalities.

Keywords

Test Collection, Dense Retrieval, Multimodal Search, Retrieval Augmented Generation, Cross Lingual Retrieval

1. Introduction

In recent years, dense retrieval technologies have seen significant advancements, largely driven by breakthroughs in large language models (LLMs). These dense retrieval methods, which use learned representations to match queries and documents, have achieved state-of-the-art performance in information retrieval tasks compared to traditional sparse approaches [7]. However, while research in this area has progressed rapidly for English-language content, non-English languages, including Japanese, have seen limited development [17]. This gap has been highlighted by the multilingual LLMs trained predominantly on English-centric datasets, which limits their effectiveness for languages with distinct linguistic characteristics, such as Japanese.

Despite the growing capabilities of multilingual models, there remains a significant lack of research and resources dedicated to developing and evaluating dense retrieval systems in non-English contexts. Particularly for Japanese, whose performance is in the bottom group in mMARCO experiment [8], there is a pressing need for dedicated research datasets that can support the development of effective ranking technologies. Such datasets (e.g., JMTEB[1]) would help address unique challenges in Japanese information retrieval, handling complex orthographic variants and providing accurate retrieval results for diverse user queries.

This paper aims to contribute to filling this gap by building test collections that are specifically designed to evaluate and advance dense retrieval technologies for multiple languages, including Japanese. By focusing on the development of resources for Japanese ranking technologies, as well as cross-lingual datasets involving Japanese, Chinese, and English, we seek to foster research that ensures these

EMTCIR '24: The First Workshop on Evaluation Methodologies, Testbeds and Community for Information Access Research, December 12, 2024, Tokyo, Japan
*Corresponding author.
†These authors contributed equally.
hideo@slis.tsukuba.ac.jp (H. Joho); a.keyaki@r.hit-u.ac.jp (A. Keyaki); tachioka.yuki@core.d-itlab.co.jp (Y. Tachioka); syamamoto@slis.tsukuba.ac.jp (S. Yamamoto)

[1]https://github.com/sbintuitions/JMTEB

technologies are effective for a broader range of users, thereby contributing to a more inclusive global information access landscape.

To address the above challenge, the NTCIR Transfer Task [1] was developed to promote research in information retrieval and natural language processing for non-English languages, with a particular focus on Japanese. The first round of the task focused on dense retrieval and reranking for Japanese, addressing challenges related to retrieval effectiveness of adhoc retrieval task. The second round expands significantly by incorporating new subtasks, including retrieval augmented generation task, cross-lingual retrieval with Japanese, Chinese, and English, and sensory data retrieval that is not language-specific. These additions aim to cover more diverse document types and data modalities, encouraging participants to develop technologies for handling both language-specific and multimodal retrieval challenges. By organizing these subtasks in a structured manner, the NTCIR Transfer Task facilitates a shared platform for evaluating dense retrieval technologies across different contexts, ultimately contributing to a more equitable and effective information retrieval landscape.

Latest information about the development of NTCIR Transfer-2 Task can be obtained from the website[2].

2. Transfer-1

The first round of the NTCIR Transfer Task, called Transfer-1, was a pilot task held at NTCIR-17 [4]. 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.

In the Dense First-Stage Retrieval subtask, participants were required to retrieve relevant documents from a large collection using dense representations. This subtask was essentially an ad-hoc retrieval task where participants were asked to use the `title` field of the original topic files as input for both the training and evaluation sets. A sample query, for instance, was *feature dimensionality reduction* (`qid:0005` in the `train` set), and the expected output was the top 1,000 document IDs. The goal was to assess how well dense retrieval methods, using semantic embeddings, could match queries to relevant documents. The evaluation metric used was nDCG@1000 (Normalized Discounted Cumulative Gain), which measured the ranking quality with a binary relevance judgment.

The Dense Reranking subtask aimed to develop second-stage retrieval techniques in a multi-stage retrieval framework. Specifically, participants were tasked with reranking the top 1,000 documents that were initially retrieved by a BM25 model. The input for this subtask was the query and the top 1,000 document IDs, and the output was a reranked list of the top 100 document IDs. To achieve this, participants often leveraged more computationally intensive models such as BERT to refine the ranking based on a deeper contextual understanding of the documents. Notably, the documents were retrieved using PyTerrier [5] (v 0.9.2), with both queries and documents tokenized by SudachiPy [6] (v 0.5.4) using its core dictionary and `SplitMode.A`. The evaluation metrics for this subtask included nDCG@20 and MRR (Mean Reciprocal Rank), focusing on both retrieval accuracy and reranking effectiveness.

Transfer-1 used the ad-hoc retrieval test collections developed at NTCIR-1 [2] and NTCIR-2 [3] as the training (`train`) set and evaluation (`eval`) set, respectively. The `train` set consisted of over 330K documents with 83 search topics, while the `eval` set consisted of 735K documents with 49 topics. The documents in the training set included titles and abstracts of academic conference papers (1988-1997), while those in the evaluation set were titles and abstracts of academic conference papers (1997-1999) and grant reports (1988-1997). It is important to note that the evaluation set included documents from the training set, but the topics and relevance judgments were independent of each other.

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 Japanese information retrieval.

[2] https://github.com/ntcirtransfer/transfer2/discussions

3. Transfer-2

Transfer-2, the second round of the task sets three subtasks: Retrieval Augmented Generation (RAG), Dense Multimodal Retrieval (DMR), and Dense Cross-Language Retrieval (DCLR).

3.1. 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 retrieval module of RAG [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 design of prompts ideal for RAG, such as the number of documents given to the generator, is also a focus.

There are two major issues to consider in conducting this subtask: 1. The design of input/output and evaluation, and 2. the design of a feasible development environment for participants. Regarding 1, fixing the query and searching target documents is necessary to ensure a fair comparison. Also, careful consideration is needed to find a method for evaluating LLMs that allows for reproducibility. Regarding 2, it is necessary to use a practical-level LLM to generate responses that withstand verification. However, the problem is that operating such a high-performance LLM requires owning a high-performance GPU.

Based on these considerations, we have decided to focus on the open-domain factoid question answering task for the RAG subtask because facts are presented in response to questions in the open-domain factoid question answering task, making evaluation easier. Additionally, we have adopted a two-stage retrieval model for the open-domain factoid question answering task, specifically the retriever-reader model used in open-domain question answering 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 either classical sparse vector search models like 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 Fusion-in-Decoder (FiD)[3] [16], a model specialized for open-domain question answering tasks. Note that FiD is not as large as cutting-edge large language models, making it a model that we believe is accessible to many participants.

We believe we can achieve our objectives by evaluating the performance and analyzing the trends of both the retriever and reader. Furthermore, we will use the AIO Official Dataset Version 2.0[4] 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 questions for training, 1,000 questions for development, and 1,000 questions for testing. Furthermore, the target corpus consists of Wikipedia articles[5].

Next, we will explain the evaluation method. The input for the first stage is a natural language question, and the output is the top 100 ranked passage IDs corresponding to the natural language question. We will use the training and development data from the AIO Official Dataset Version 2.0 for training and evaluation data, respectively. The evaluation metrics will be HitRate@k ($k = 1, 5, 10, 50, 100$) and

[3] https://github.com/facebookresearch/FiD
[4] https://sites.google.com/view/project-aio/dataset
[5] https://github.com/cl-tohoku/quiz-datasets

nDCG@k ($k = 1, 5, 10, 50, 100$). For the second stage, the input is a natural language question and k passages retrieved in the first stage, and the output is the answer to the natural language question. The evaluation metric will be accuracy.

3.2. DMR subtask

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 determine positive and hard negative samples for contrastive learning [10], particularly due to lack of established datasets paired with non-linguistic modalities. Despite the advancements in LLMs 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 remain unclear [11]. While large text and image corpora exist (e.g., Microsoft COCO [12] and Flickr30k [13]), they often lack 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.

To effectively advance this subtask, we intend to 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 are currently building 30 topics for the validation and test. Each topics is generated on a daily basis, and sensors or images recorded at the same time as the query are extracted as relevant data. Additionally, we design two types of retrieve tasks: one that retrieves images from sensors and another that retrieves sensors from images.

3.3. DCLR subtask

This subtask aims to build test collections that can be used to study cross-lingual information retrieval and multilingual information retrieval with a focus on dense Japanese technologies. The retrieval task is compatible to Transfer-1 including the first stage retrieval and reranking for adhoc retrieval.

Thanks to the organizer of TREC NeuCLIR Track, this subtask will provide corpora in up to three languages such as Japanese, English, and Chinese, reflecting societal interests in Asian regions. Of those, both Japanese and English corpora will be newly generated for this subtask based on the protocol developed by the NeuCLIR Track. We plan to use an existing corpus for Chinese from the same Track. The corpora will be news-related documents from the CommonCrawl collection dating from 2016 to 2021. The number of documents will be in the range of two to three millions. We plan to release these corpora via HuggingFace.

We are currently building 50 new topics for the adhoc retrieval task including the title, description, and narratives. The original topic language will be Japanese, and we plan to manually translate them into English and Japanese.

As for relevance assessments of retrieved documents, we plan to apply a hybrid approach of LLM-based judgements with few shot human annotations. We have identified three to five relevant items during the topic creation process. These relevant items can be used in the prompt design of relevance assessments by LLM. We plan to compare the outcome of relevance assessments between a proprietary model such as gpt-4o mini and open source model such as Llama 3. Manual validation on a stratified

sampling (relevant and non-relevant) will be conducted to demonstrate the accuracy of LLM-based relevance assessments. Although the reliability of LLM-based relevance assessments has not been established yet, we aim to explore the utility of a hybrid approach towards this direction.

4. Conclusive Discussion

The development of dense retrieval technologies for non-English languages, particularly Japanese, remains a challenging but crucial area for advancing multilingual information access. The NTCIR Transfer Task, including its first and second rounds, aim to contribute to addressing this gap by building specialized test collections and organizing focused tasks. These efforts encourage researchers to explore effective dense retrieval, cross-lingual information retrieval, and multimodal search technologies. By fostering the development of resources like dedicated datasets for Japanese and other languages, we strive to create a more inclusive and effective global information retrieval landscape.

The results and insights gathered through the NTCIR Transfer-1 Tasks indicated that dense retrieval methods are promising but need further enhancement in handling specific language characteristics and diverse modalities. Transfer-2, with its emphasis on Retrieval-Augmented Generation, Dense Multimodal Retrieval, and Dense Cross-Language Retrieval, lays the foundation for the next phase of research and provides valuable resources and benchmarks for the research community.

Future work will focus on extending these test collections, refining dense retrieval methods to handle diverse languages and modalities effectively, and evaluating the cross-lingual models to ensure a fair representation for different languages. We anticipate that the continued efforts will contribute to advancements not only in dense retrieval but also in promoting equitable information access across languages and communities.

Acknowledgments

The authors thank to the general chairs and program chairs of NTCIR-18, and NTCIR Office for their support to run Transfer Task. The authors also thank the anonymous reviewers for their constructive comments on an earlier version of the manuscript. The research was partly supported by ROIS NII Open Collaborative Research 2024 (Grant Number: 24S0503). The opinions, findings, and conclusions presented in this paper are those of the authors and do not necessarily reflect the views of the sponsors.

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] Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato, and Souichiro Hidaka. (1999). Overview of IR Tasks at the First NTCIR Workshop. In: Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, August 30 - September 1, 1999, pp.11-44.

[3] Noriko Kando, Kazuko Kuriyama, Masaharu Yoshioka. (2001). Overview of Japanese and English Information Retrieval Tasks (JEIR) at the Second NTCIR Workshop. In: Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization, May 2000- March 2001. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings2/ovview-kando2.pdf

[4] 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

[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.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3733–3742. CVPR'18 (2018)

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

[12] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: 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. Springer (2014)

[13] Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: 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.: 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 (2024)

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

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

[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.