projects and focus more on projects which gather data and edit pages. The DH community can learn from the technical community regarding the factors that influence its data quality, and possible solutions. Factors specified in the research include: user types and their editing activities, the effectiveness of systems and tools to facilitate detection and improvement of data quality, and the relevance and authoritativeness of its external references and sources. The solutions proposed by the technical community encompass 1) a better understanding of users and the editorial process via research, and 2) the development of systems, measures, and tools concerning the evaluation and improvement of different dimensions of data quality. The technical side, however, has its limitation. As pointed out by the IS systematic review (Mora-Cantallops et al., 2019, 262), such applications are mostly limited to Wikidata itself and are yet to be linked to disciplines outside information systems. The contribution of this paper is to address three factors and relevant solutions in the specific context of DH projects: the relevance and authoritativeness of other available domain sources, domain communities and their activities, and workflow designs that balance the automated and manual work by utilising the technical and labour resources of a project's own and those offered by Wikidata.

This paper intends to invite discussion from participants at DH2022 about Wikidata's possible use in the DH context and the challenges it may face.

Bibliography

Cook, S. (2017). The uses of Wikidata for galleries, libraries, archives and museums and its place in the digital humanities. *Comma*, **2017**(2):117-124.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, **33**(2004), 1-26.

Mora-Cantallops, M., Sánchez-Alonso, S. and García-Barriocanal, E. (2019). A systematic literature review on Wikidata. *Data Technologies and Applications*, **53**(3): 250–68.

Tharani, K. (2021). Much more than a mere technology: A systematic review of Wikidata in libraries. *The Journal of Academic Librarianship*, **47**(2).

Notes

1. Data about the ADHO annual conferences is collected from *the Index of Digital Humanities Conferences* site which aggregates and presents conference metadata: https://dh-abstracts.library.cmu.edu/conferences
2. Until December 31, 2021.

Multimedia Retrieval of Historical Materials

Zhu, Jieyong
zjsczjy04@gmail.com
Graduate School of Informatics, Kyoto University

Nishimura, Taichi
taichitary@gmail.com
Graduate School of Informatics, Kyoto University

Goto, Makoto
m-goto@rekihaku.ac.jp
National Museum of Japanese History

Mori, Shinsuke
forest@i.kyoto-u.ac.jp
Academic Center for Computing and Media Studies, Kyoto University

Introduction

Historical material is a collection of history, archaeology, and folklore materials. With the rapid advancement of digitization, large-scale multimedia data of historical materials have become available on the web. As the data grows, it is difficult for researchers to study the relationship between historical images and text. Multimedia retrieval is a technique to perform retrieval tasks across multiple media. Recently, deep learning has accelerated research on natural language understanding and computer vision, with remarkable performance reported in multimedia retrieval tasks (Salvador et al., 2017). In this paper, we apply the state-of-the-art multimedia retrieval methods to Japanese historical materials and demonstrate the constructed multimedia retrieval system. Fig 1 shows an example of multimodal retrieval tasks of historical materials.

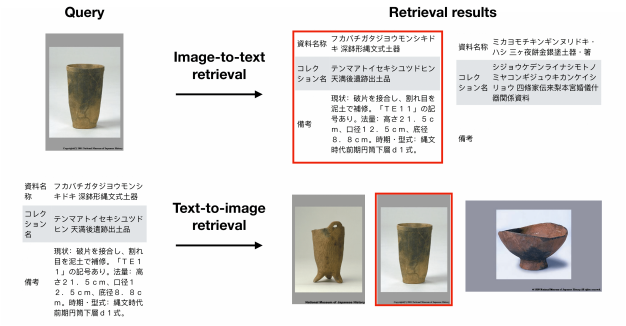

Fig 1.
Examples of image-to-text and text-to-image retrieval tasks. Retrieved results in the boxes are the associated ones with the query on the left.

Multimedia Retrieval

Multimedia retrieval takes one type of media (e.g., images and texts) as the query to retrieve corresponding media of another type (Liu et al., 2010). The key challenge of multimedia retrieval is how to convert different media data into a shared subspace, where semantically associated inputs are mapped to similar locations. Various kinds of deep-learning-based approaches have been proposed in the literature (Sirirattanapol et al., 2017). We here employ one of them to realize our system.

Proposal

This paper proposes a deep-learning-based approach to achieve multimedia retrieval of historical materials. Figure 2 shows an overview of our proposed model. The proposed method consists of two major processes.

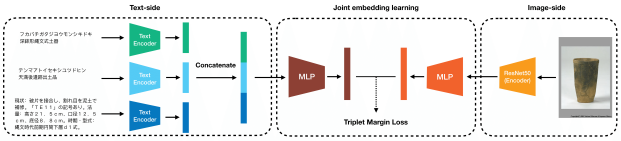

Fig 2.
An overview of our proposed method.

Text encoder

Recently, large-scale pre-trained model, such as BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), has achieved great performance in NLP tasks. However, we don't use BERT model because of the domain gap, since Japanese BERT is trained on Japanese Wikipedia texts. The text data of historical materials include three main types: material name, collection name, and notes. All the three types of data are in a tabular format instead of a sentence format. Therefore, we use a word2vec model to convert the texts into vectors, which is more simple and more reasonable.

Image encoder

The input of the image side is a single image of historical materials. To convert images into vectors, we employ ResNet50 (He et al., 2016), a Convolutional Neural Network pre-trained on ImageNet.

Shared subspace learning

Finally, we convert text/image vectors into shared subspace using symmetric multi-layer perceptrons with ReLU activation functions. To train the model, we compute triplet margin loss (Vassileios Balntas and Mikolajczyk, 2016), which makes the vectors in the subspace for a given text-image pair close and otherwise long.

Dataset

This is the first attempt to tackle multimedia retrieval of historical materials, so no datasets exist in this field; thus we created the Japanese historical dataset of textual descriptions and corresponding images by crawling them from the National Museum of Japanese History. The dataset contains 18,429 objects, including over 18k textual descriptions and over 79k corresponding images.

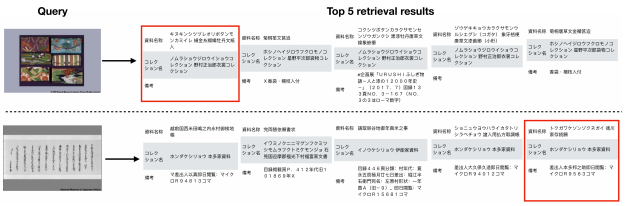

Fig 3.
Image-to-text retrieval examples. The ground truth in the retrieved results is highlighted in the box.

Experiments

To measure the performance of the model, we perform multimedia retrieval tasks. Figure 3 shows two examples of the image-to-text task. The query images are on the left side while the top five retrieved texts are on the right side. As with previous studies, we compute three mainstream

evaluation metrics, median rank (MedR), Recall@K (R@K) (Salvador et al., 2017), and mean average precision (mAP) (Rasiwasia et al., 2010) to evaluate the performance. Table 1 shows the results of 1,000 samples. The result indicates that our system performs well in multimedia retrieval tasks compared with the random ranking baseline.

	Image => Text	Text => Image	Random Ranking
R@1	0.036	0.043	0.001
R@5	0.144	0.163	0.005
R@10	0.258	0.285	0.01
medR	26	26	500
mAP	0.107	0.119	0.002

Table 1.
Retrieval results on 1,000 samples.

Conclusion

This paper tackled the multimedia retrieval of historical materials using deep-learning-based multimedia retrieval methods. This work is the first attempt to tackle this problem, thus we constructed the dataset of Japanese historical texts and images, and evaluated the model's performance on it. The experimental results show that our constructed system performs well in the multimedia retrieval of historical materials. Future work will study a better method to represent the textual data. We expect that our research will help researchers in gaining a better understanding of Japanese historical materials, and will give a general approach to learning the shared subspace between textual and visual data.

Bibliography

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–86 doi:10.18653/v1/N19-1423. https://aclanthology.org/N19-1423.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 770–78.

Liu, J., Xu, C. and Lu, H. (2010). Cross-media retrieval: state-of-the-art and open issues. *International Journal of Multimedia Intelligence and Security*, **1**(1). Inderscience Publishers: 33–52.

Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G. R., Levy, R. and Vasconcelos, N. (2010). A new approach to cross-modal multimedia retrieval. *Proceedings of the 18th ACM International Conference on Multimedia*. pp. 251–60.

Salvador, A., Hynes, N., Aytar, Y., Marin, J., Ofli, F., Weber, I. and Torralba, A. (2017). Learning cross-modal embeddings for cooking recipes and food images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 3020–28.

Sirirattanapol, C., Matsui, Y., Satoh, S., Matsuda, K. and Yamamoto, K. (2017). Deep image retrieval applied on kotenseki ancient japanese literature. *2017 IEEE International Symposium on Multimedia (ISM)*. IEEE, pp. 495–99.

Vassileios Balntas, D. P., Edgar Riba and Mikolajczyk, K. (2016). Learning local feature descriptors with triplets and shallow convolutional neural networks. In Richard C. Wilson, E. R. H. and Smith, W. A. P. (eds), *Proceedings of the British Machine Vision Conference (BMVC)*. BMVA Press, p. 119.1-119.11 doi:10.5244/C.30.119. https://dx.doi.org/10.5244/C.30.119.