Researcher Representations Based on Aggregating Embeddings of Publication Titles: A Case Study in a Japanese Academic Database

Hiroyoshi Nagao
Doshisha University
nagao21@mm.doshisha.ac.jp

Marie Katsurai
Doshisha University
katsurai@mm.doshisha.ac.jp

Abstract

Constructing researcher representations is crucial for search and recommendation in academic databases. While recent studies presented methods based on knowledge graph embeddings, obtaining a complete graph of academic entities might be sometimes challenging due to the lack of linked data. By contrast, the textual list of publications of each researcher, which represents their research interests and expertise, is usually easy to obtain. Therefore, this study focuses on creating researcher representations based on textual embeddings of their publication titles and assesses their practicality. We aggregate embeddings of each researcher's multiple publications into a single vector and apply it to research field classification and similar researcher search tasks. We experimented with multiple language models and embedding aggregation methods to compare their performance. From the model perspective, we confirmed the effectiveness of using sentence embedding models and a simple averaging approach.

1 Introduction

Academic databases equipped with search and recommendation functions have supported researchers' activities. Traditional search approaches primarily involve users inputting specific keywords and then retrieving corresponding papers and authors (Wu et al., 2019; Fricke, 2018). However, in actual research scenarios, researchers often need to efficiently find papers to read based on their previous interests or discover individuals in other fields with similar research interests. To address these requirements, methods for constructing researcher representations in a feature space have been actively studied in recent years (Chaiwanarom and Lursinsap, 2015; Katsurai et al., 2016; Ganesh et al., 2016; Färber et al., 2023). In particular, with the development of academic graphs such as Microsoft Academic Graph (MAG)[1], there have been presented several methods based on knowledge graph embeddings to construct researchers' vector representations (Priem et al., 2022; Färber et al., 2023). However, these studies assume well-organized graph databases as input resources, making it difficult to apply these methods to databases lacking sufficient linked data. For example, Grant-in-Aid for Scientific Research database (KAKEN), one of typical academic databases in Japan, comprises records of research accomplishments of grant projects but does not link them to their authors and presentation venues. Using such databases would first require data linking to construct a graph structure before applying graph embedding methods.

Compared with obtaining a complete graph structure, the textual list of publications of each researcher is more accessible. Recent development of language models pre-trained using large-scale text data (Devlin et al., 2019; Feng et al., 2022; Wang et al., 2022) has significantly improved several retrieval and recommendation tasks, and academic-specific language models were released for certain languages (Beltagy et al., 2019; Labrak et al., 2023; Yamauchi et al., 2022). Under such background, this study attempts to construct researcher representations solely from the textual publication information using pre-trained language models. The usefulness of these representations is then examined through experiments on two practical application tasks: research field classification and similar researcher search.

2 Related Work

Researcher representations have been studied using a variety of methods, such as topic models (Chaiwanarom and Lursinsap, 2015; Katsurai et al., 2016)

[1] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/

Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024), pages 277–282

and document embeddings (Ganesh et al., 2016). Recent methods focused on the graph structure of academic entities and calculated the node features using graph embedding methods. MAG was a typical example of a large-scale academic graph, and its successor project, OpenAlex (Priem et al., 2022), has recently been used. Färber et al. (2023) constructed an RDF-based knowledge graph platform called SemOpenAlex and experimentally evaluated the embedding methods for various nodes such as researchers and papers on the graph. Such knowledge graph-derived researcher representations are used to solve academic tasks such as author name disambiguation (Santini et al., 2022).

On the other hand, the newly emerging pre-trained language models have substantially promoted the development of text processing for information retrieval and recommendation, and we can expect their effectiveness on researcher embeddings. It is known that a certain domain task often benefits from a language model pre-trained using text in the target domain. However, in scientific domains, we observed cases where general domain models are sufficient (Zheng et al., 2022; Brokman and Kavuluru, 2024). Taking into account the above matters, this paper includes an experiment that compares the performance between general language models and domain-specific models in a single language embedding setting.

3 Methodology

Our method aims to construct a researcher's representation using a comprehensive list of publication titles related to their research grant project. Since most researchers in Japan use both English and Japanese languages to present their research, we investigate in experiments whether we should embed both the two languages' text to calculate the researchers' representations.

3.1 Text Embedding Based on Pre-trained Language Models

To acquire semantic information from the text of each publication title, we exploit the BERT architecture (Devlin et al., 2019) as a pre-trained text encoder. First, the given text is divided into subwords by the tokenizer associated with the model. These subwords are then input into the model to obtain embedding representations at the subword level. Next, pooling operations are applied to the obtained token representations to derive the publication text embeddings. There are two choices for pooling: CLS pooling, which uses the representation of the special initial token, and mean pooling, which averages the representations of all tokens. Based on a conventional study (Reimers and Gurevych, 2019), we use mean pooling.[2]

Language	Count
Japanese	3,122,244
English	3,283,871
Others	779

Table 1: Statistics of languages used in KAKEN projects.

3.2 Aggregating Publication Text Embeddings

The interests of researchers are diverse, and it is not clear how to best aggregate publication titles' embeddings. The first method is arithmetic averaging (hereafter referred to simply as **averaging**), commonly used for pooling word representations into sentence representations. However, researchers do not necessarily contribute equally to all their publications. For example, contributions are generally more distributed in co-authored papers compared to single-authored papers. Therefore, as the second method, we consider **weighted averaging** using the inverse of the number of authors as weights. This assumes that the fewer the authors, the greater the individual contribution to the publication. This method is inspired by author evaluation metrics like the h-index, which takes into account the influence of the number of authors (Abramo et al., 2013). In subsequent experiments, we compare the performance of these two methods.

4 Experiments

We used the KAKEN API[3] to collect KAKEN projects as of October 2023 and the publication titles linked to the projects' members. We explain details about the KAKEN API in Appendix A. Table 1 shows the distribution of languages in the publication titles. As shown in Table 1, Japanese researchers usually publish papers in Japanese and English. "Other" category included languages such as Russian and Mongolian. We exploited multilingual models to embed Japanese and English

[2]Actually, in downstream tasks mentioned later, mean pooling outperformed CLS pooling in all tasks.

[3]https://support.nii.ac.jp/en/kaken/api/api_outline

Category	Research field name	Count
A	arts	26,170
B	algebra	5,812
C	strength of materials	7,451
D	materials engineering	4,477
E	organic/inorganic chemistry	3,295
F	agriculture	5,372
G	biology	3,177
H	pharmacy	3,876
I	internal/social medicine	27,165
J	information science	3,591
K	environmental analysis	1,675

Table 2: Large categories of research fields in KAKEN.

Model	Large F1	Medium F1	Small F1
mE5 (avg)	0.606	**0.514**	**0.396**
mE5 (weighted)	0.604	0.510	0.361
mBERT	0.582	0.455	0.311
LaBSE	**0.607**	0.504	0.367

Table 3: Results of research field classification.

publication titles into the same space. Specifically, we adopted the following three models: Multilingual BERT (mBERT)[4] (Devlin et al., 2019), Multilingual E5 (mE5)[5] (Wang et al., 2022), and LaBSE[6] (Feng et al., 2022); each model had 110 million parameters. Among these, mE5 and LaBSE are so called sentence embedding models. We aim to find out whether such models are suitable for embedding publication titles.

To verify the effectiveness of the constructed researcher representations, we conducted two practical application tasks: research field classification (see Section 4.1) and similar researcher search (see Section 4.2). Since there is no multilingual pretrained model in the scientific domain, we also investigate the performance of domain-specific models in single language embedding setting (see Section 4.3).

[4] https://huggingface.co/bert-base-multilingual-cased

[5] https://huggingface.co/intfloat/multilingual-e5-base

[6] https://huggingface.co/sentence-transformers/LaBSE

4.1 Research Field Classification

Research field classification is a task to identify the main research fields of a researcher based on their representation. The ground-truth labels were determined based on area categories[7] of the research projects associated with each researcher. Each research project was manually classified by its representative researcher in terms of three levels: large, medium, and small categories. These levels have 11, 65, and 305 categories, respectively. Table 2 shows area category names and number of researchers for only large categories due to space constraints. We provided three-level labels for each researcher using their project labels. Appendix B describes the labeling procedure. We obtained a dataset of 92,061 researchers who have at least three publication titles. All data was split into training, validation, and test sets in an 8:1:1 ratio, keeping the original label distribution. At the training stage, we added a linear layer as a classifier. Specifically, we froze the weights of the text encoder and trained only the linear layer. We used a scheduler that increases the learning rate until a certain point and then fixes it. We set the period of time until the learning rate was fixed to one epoch.

Table 3 shows the F1 scores of classification results. In Appendix B, we also evaluate the accuracy. The model "avg" denotes aggregation using the averaging, while "weighted" represents aggregation using the weighted averaging. Notably, aggregation using the averaging exhibited superior performance, indicating that weighting by the inverse of the number of authors did not enhance performance. Hereafter, references to each model refer to the case where aggregation using the averaging was employed. Focusing on different models, mE5 and LaBSE demonstrated similar performance. Conversely, mBERT exhibited a significant performance gap compared to mE5, with 5% lower F1 scores in the middle sections and 8% lower in the small sections, although no significant performance difference was observed in the large category classification.

4.2 Similar Researcher Search

This experiment defined positive and negative examples for a query researcher. We ranked the com-

[7] Because category names are reviewed and updated at a few years intervals, we focused on the category hierarchy introduced in 2018 fiscal year. Details are available in https://www.jsps.go.jp/j-grantsinaid/02_koubo/shinsakubun.html.

Model	MAP	NDCG
mE5 (avg)	**0.783**	**0.856**
mE5 (weighted)	0.758	0.839
mBERT	0.711	0.805
LaBSE	0.765	0.843

Table 4: Results of researcher search based on research content similarity.

Language	Model	MAP	NDCG
Multi	mE5	**0.783**	**0.856**
English	mE5	0.740	0.815
	SciBERT	0.701	0.785
	BERT	0.650	0.747
Japanese	mE5	0.777	0.851
	AcRoBERTa	0.576	0.696
	RoBERTa	0.514	0.645

Table 5: Comparison among different language resources in the researcher search task.

bined set of researchers using the similarity between the researcher representations and evaluated whether the positive examples were ranked higher. For each query, we defined co-investigators of research projects as positive examples, and randomly selected other researchers as negative examples. From the 92,061 researchers, we extracted up to 10 queries per small research category, totaling 2,980 queries, where each query had at least 5 positive examples. We used cosine similarity as the measure of similarity between the researcher representations. We set the maximum number of positive examples to 10, and if there were more than 10 positive examples, we preferentially selected those associated with more research projects. We set the number of negative examples to 50.

Table 4 shows the MAP and NDCG of results obtained by each model's similarity measure. As shown, mE5 achieved the best performance, followed by LaBSE and mBERT. Similar to research field classification, there was no significant performance disparity between mE5 and LaBSE, while a noticeable 7% difference in MAP was noted compared to mBERT. Combined with the results in section 4.1, we observed that the sentence embedding models positively influenced the embedding of paper titles and their subsequent aggregation.

4.3 Influence of Input Language

In the researcher search task, we further investigated the influence of multilingualism and domain adaptation, which can be used as future guidelines for designing a multilingual model specialized in the academic domain. We considered the following three settings: (i) embedding only Japanese titles, (ii) embedding only English titles, and (iii) embedding both Japanese and English titles. All settings used mE5 as a multilingual model, which was the best performing model in the previous experiments. The first setting used Academic RoBERTa (Yamauchi et al., 2022) (hereafter, **AcRoBERTa**), a model specific to the Japanese academic domain, and normal RoBERTa[8]. The second setting used SciBERT (Beltagy et al., 2019), a model specific to the English academic domain, and normal BERT. The third setting used only mE5 and no additional models.

Table 5 shows the results of different models for each setting. Focusing on mE5, using multilingual inputs resulted in higher performance than using only Japanese or only English inputs. This suggests that better feature representations are obtained by aggregating embeddings of publications in both Japanese and English titles of researchers. On the other hand, comparing the results for only English inputs, mE5, which is not specialized, showed higher performance than the domain-specific SciBERT. This observation was also noted when the input was only Japanese. Since AcRoBERTa has a very small training corpus compared to other models, it is considered that models trained on large corpora are advantageous for obtaining researcher representations in all fields. In conclusion, at least for inputs such as publication titles that are relatively short texts, it is suggested that the influence of adaptation to sentence embeddings is more important than the effect of domain adaptation.

5 Conclusion

In this paper, we focused on the titles of researchers' publications in academic databases and constructed researcher representations by aggregating their text embeddings. We then evaluated the performance of these representations in two practical tasks. The results of the experiments showed that the performance of each task improved with sentence embedding models. Additionally, inputting publica-

[8]https://huggingface.co/nlp-waseda/roberta-base-japanese

tions in multiple languages was more effective than in a single language. However, there were practical challenges with domain-specific models, as general-domain models outperformed them. These results suggest the need for the development of new multilingual models that are robust in sentence representation and have domain knowledge.

Acknowledgments

This research was partly supported by JSPS KAKENHI Grant Number JP20H04484.

References

Giovanni Abramo, Ciriaco Andrea D'Angelo, and Fulvio Viel. 2013. Assessing the accuracy of the h-and g-indexes for measuring researchers' productivity. *Journal of the American Society for Information Science and Technology*, 64(6):1224–1234.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Aviv Brokman and Ramakanth Kavuluru. 2024. How important is domain specificity in language models and instruction finetuning for biomedical relation extraction? *arXiv preprint arXiv:2402.13470*.

Paweena Chaiwanarom and Chidchanok Lursinsap. 2015. Collaborator recommendation in interdisciplinary computer science using degrees of collaborative forces, temporal evolution of research interest, and comparative seniority status. *Knowledge-Based Systems*, 75:161–172.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Michael Färber, David Lamprecht, Johan Krause, Linn Aung, and Peter Haase. 2023. Semopenalex: The scientific landscape in 26 billion rdf triples. In *The Semantic Web – ISWC 2023*, pages 94–112, Cham. Springer Nature Switzerland.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Suzanne Fricke. 2018. Semantic scholar. *Journal of the Medical Library Association: JMLA*, 106(1):145.

J Ganesh, Soumyajit Ganguly, Manish Gupta, Vasudeva Varma, and Vikram Pudi. 2016. Author2vec: Learning author representations by combining content and link information. In *WWW (Companion volume)*, pages 49–50.

Marie Katsurai, Ikki Ohmukai, and Hideaki Takeda. 2016. Topic representation of researchers' interests in a large-scale academic database and its application to author disambiguation. *IEICE Transactions on Information and Systems*, 99(4):1010–1018.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. DrBERT: A robust pre-trained model in French for biomedical and clinical domains. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16207–16221, Toronto, Canada. Association for Computational Linguistics.

Jason Priem, Heather Piwowar, and Richard Orr. 2022. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Cristian Santini, Genet Asefa Gesese, Silvio Peroni, Aldo Gangemi, Harald Sack, and Mehwish Alam. 2022. A knowledge graph embeddings based approach for author name disambiguation using literals. *Scientometrics*, 127(8):4887–4912.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Jian Wu, Kunho Kim, and C. Lee Giles. 2019. Citeseerx: 20 years of service to scholarly big data. In *Proceedings of the Conference on Artificial Intelligence for Data Discovery and Reuse*, AIDR '19, New York, NY, USA. Association for Computing Machinery.

Hiroki Yamauchi, Tomoyuki Kajiwara, Marie Katsurai, Ikki Ohmukai, and Takashi Ninomiya. 2022. A Japanese masked language model for academic domain. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 152–157.

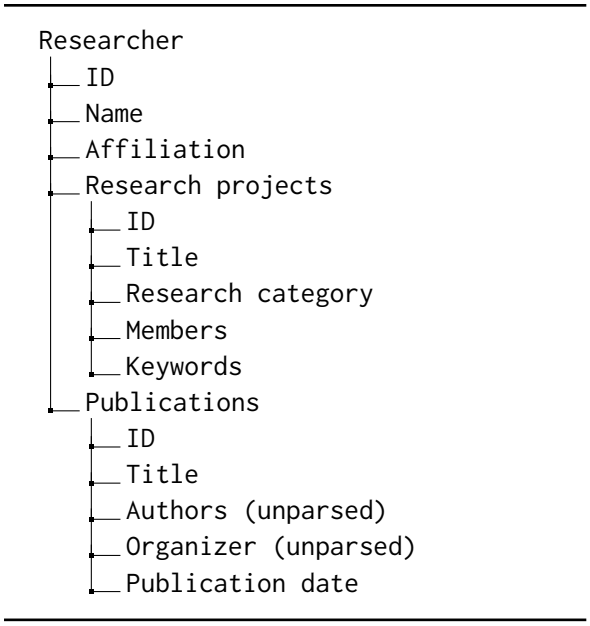

Figure 1: Researcher data that can be obtained using the KAKEN API.

Zhe Zheng, Xin-Zheng Lu, Ke-Yin Chen, Yu-Cheng Zhou, and Jia-Rui Lin. 2022. Pretrained domain-specific language model for general information retrieval tasks in the aec domain. *arXiv preprint arXiv:2203.04729*.

A Use of KAKEN API

KAKEN is a public database that includes information regarding projects supported by KAKENHI, a grant program provided by the Japan Society for the Promotion of Science. The program supports research projects in diverse research fields, including natural sciences, engineering, social sciences, and humanities, for researchers affiliated with Japanese institutions. KAKEN provides an API that allows a search of KAKENHI projects using several types of queries (e.g., research periods, researcher names). Figure 1 shows items that can be obtained for each researcher through the KAKEN API.

B Labeling Procedure for Research Field Classification

We determined a unique field in each category level for a researcher using the following method. We first extracted the research projects with explicitly defined small categories from the set of research projects held by the researcher, and assigned the most frequent small category as the ground-truth one. Next, we listed the large, medium pairs to which the small research category belongs in the category hierachy defined in KAKEN. If the pair was uniquely determined, the process was completed. Finally, for each listed pair, we calculated the sum of the frequencies of the large and medium sections from the set of research projects. The pair with the highest frequency was then determined.

Model	Large Acc	Medium Acc	Small Acc
mE5 (avg)	**0.759**	**0.588**	**0.475**
mE5 (weighted)	0.757	0.586	0.468
mBERT	0.744	0.530	0.398
LaBSE	0.756	0.580	0.473

Table 6: Results of research field classification.

C Accuracy of Research Field Classificaiton

At the training stage, we searched for the maximum learning rate in the range of 1e-5, 3e-5, 5e-5, 1e-4 and adopted the evaluation value of the trial with the highest performance on the validation set. Table 6 shows accuracy of research field classification. As described in Section 4.1, in terms of the aggregation method, the averaging was slightly higher than the weighting by the inverse of the number of authors. In the comparison between models, as with the F1 Score, a notable difference in performance was observed between mE5 and mBERT. In the comparison between the models, the difference in performance between mE5 and mBERT was 8% in the small section, which is as significant as in the F1 Score.