Scientometrics (2021) 126:2311–2327
https://doi.org/10.1007/s11192-020-03861-3

Multilingual author matching across different academic databases: a case study on KAKEN, DBLP, and PubMed

Yuto Chikazawa[1] · **Marie Katsurai**[1] · **Ikki Ohmukai**[2]

Received: 15 August 2020 / Accepted: 31 December 2020 / Published online: 5 February 2021

Abstract
Researchers often use their native languages to present and exchange ideas. To construct an individual author's complete profile, a list of their English and non-English academic publications must be constructed. This paper presents a practical approach for multilingual author matching across different academic databases. Our approach automatically links the academic records of a target database to a researcher identifier of a source database. First, we extracted a comprehensive set of records in the target database, whose author names were identical to the researcher names in the source database. Then, we calculated multiple author similarity measures, which can be adopted in certain entity pairs from different language databases. Finally, we aggregated the measures to output an improved score that indicates the likelihood of each record as being the researcher's work. Our method was found to be easy to implement, and its performance was evaluated in real database management settings. Experiments were conducted using DBLP and PubMed as the target English databases. As the Japanese database, KAKEN was the source for identifying researcher information. The results demonstrated each similarity measure's performance, from which we observed that the score aggregation achieved stable performance. Our method can lessen human efforts to associate various scholarly contributions.

Keywords Multilingual author matching · Academic databases · KAKEN · DBLP · PubMed

Yuto Chikazawa and Marie Katsurai have equal contribution.

✉ Marie Katsurai
katsurai@mm.doshisha.ac.jp

Yuto Chikazawa
chikazawa@mm.doshisha.ac.jp

Ikki Ohmukai
i2k@l.u-tokyo.ac.jp

[1] Doshisha University, 1–3 TataraMiyakodani, Kyotanabe, Kyoto 610–0394, Japan

[2] The University of Tokyo, 7–3–1 Hongo, Bunkyo-ku, Tokyo 113-8654, Japan

Introduction

The constantly increasing availability of digital data on scholarly contributions, such as academic papers, research grants, awards, and dissertations, presents numerous opportunities to analyze the structure and development of science using data mining techniques. The "science of science" subject encompasses several topics such as performance evaluations of researchers and research institutions, analysis of collaboration patterns, and visualization of career paths for researchers. Studies on this subject, in principle, require that the authors of every scholarly contribution are accurately linked to individual researchers. However, an online academic database generally contains only a single type of scholarly contribution, and personal research information is scattered across multiple databases. Automatic aggregation of an individual researcher's contributions from various databases is not usually straightforward because of the "homography" of the full names of the researchers,[1] which prevents a direct name matching-based author identification. Thus, this research focuses on constructing an algorithm for efficient author matching across different academic databases.

Although author name disambiguation has been widely studied, most conventional methods are designed to work within a single database. Nevertheless, our problem setting substantially differs from conventional scenarios, and faces the following two open challenges. (i) *Limited common metadata*. First, different databases do not always follow the same schema, and we can use only the general attributes that are commonly used in all types of contributions. Second, constructing true author pairs for each combination of different academic databases is time-intensive, which typically makes the application of supervised machine learning algorithms difficult. (ii) *Differences of languages*. Researchers often use their native languages to facilitate national scholarly exchange of ideas and disseminate new knowledge (Salager-Meyer 2014), which creates the need for linking authors of domestic contributions to those of international publications. For example, Brazilian scientists annually publish approximately 50,000 articles (as of 2007), of which approximately 60% are in Portuguese (Meneghini and Packer 2007), and 35% of Japanese papers available in Google Scholar are written in Japanese only, with neither an English title nor an English abstract (Amano et al. 2016). Therefore, to improve the comprehensiveness of the science of science study, the authors of English and non-English academic records must be matched to construct a researcher's publication list.

Considering the above two problems, this paper presents a naive, unsupervised multilingual author matching approach as a practical solution. Our method assumes that there are two databases: one database equips an author identifier (ID) system (called a *source database*), whereas the other database does not maintain any such identification system (called a *target database*). These two databases use languages different from one another (i.e., English and non-English languages). Given a certain full name and a set of author IDs that have the name in a source database, we first extract a comprehensive set of records whose authors have the same name in the target database. Then, we present several types of similarity measures that can be calculated even for a pair of different languages. Finally, we fuse the measures to obtain a final similarity score between an author in the source database and records in the target database in an unsupervised manner. The resulting ranking

[1] Throughout the paper, following the definition in (Müller et al. 2017), we describe a group of distinct authors having the same name as *homography*.

can lessen human efforts to associate various scholarly contributions and is thus a practical solution for academic database management.

A preliminary version of this paper has been published previously in Katsurai and Ohmukai (2019). The major difference between this paper and the previous version is that we extended the target scenario from monolingual to multilingual. To evaluate the performance of the proposed method, we conducted experiments that link multiple records of two English databases, namely, DBLP[2] and PubMed,[3] to the author IDs of a Japanese grant database, namely, KAKEN.[4] The results demonstrated that the fused similarity outperformed single similarity measures.

The main contributions of this research can be summarized as follows:

- To the best of our knowledge, our work is the first to study multilingual author matching across different academic databases.
- Our method exploits the attributes that are usually available in any type of scholarly contribution and is easy to implement for practical use.
- We present a case study of profiling Japanese researchers, demonstrating that the aggregated ranking can produce stable results.

The remainder of this paper is organized as follows. The next section briefly discusses some conventional studies on unsupervised author name disambiguation, as well as the open challenges in academic database management. Then, the proposed method and details of the datasets used in this study are presented. Subsequently, the experimental results are described. Finally, in the last section, the paper is summarized, and some future research directions are suggested.

Related work

Author name disambiguation in academic databases has been actively studied as one of the essential techniques for digital library management (Ferreira et al. 2012). This section briefly describes the literature available on unsupervised author identification. Academic records usually contain attributes of authors, publication dates, and titles. Most conventional methods manually construct a similarity function suitable for each attribute type (Bhattacharya and Getoor 2007; Cota et al. 2010; Han et al. 2005), in which string similarity is often utilized. For example, Bhattacharya and Getoor (2007) calculated the similarities between attribute strings using the Jaro, Levenshtein, and Jaro–Winkler distances. They also presented a coauthor-based similarity based on the union operation, Jaccard Coefficient, and the Adamic/Adar score. Cota et al. (2010) used edit distances to calculate the string similarity between author and coauthor names. They also calculated the cosine similarities between TF-IDF features calculated from publication venues. These similarity-based methods are relatively easy to implement, which is an important factor for real digital library management. Following these works, we propose an author-matching method based on multiple similarity functions. Our measures are based on a small set of

[2] https://dblp.uni-trier.de/.
[3] https://pubmed.ncbi.nlm.nih.gov/.
[4] https://kaken.nii.ac.jp/.

metadata that is generally available, which makes it applicable to all types of scholarly contributions.

Compared with the conventional single-language and single-type scenario, matching of authors across different languages or different types of scholarly contributions has not been studied extensively. In more general settings beyond academic libraries, some works have investigated multi-database or multilingual data linking. For example, Long and Jung (2015) proposed a social identity matching method across multiple social networking sites. To calculate the likelihood of whether two users are the same individual, they used username string similarity and the users' social relationships, which were easy to obtain under the policies of the social networking sites. When considering the application of this conventional method to our problem, we found that its similarity calculation module has room for improvement because textual data (especially reflecting the researcher's interests) can also be powerful features in academic libraries. In multilingual contexts, Jung (2013) focused on the use of several languages by social media users for tagging. They presented a tag-matching method across different languages based on the co-occurrence frequency of tags assigned by multilingual speakers. Gupta et al. (2014) highlighted the difficulty in searching a database that contains transliterations, which are converted from original languages. To comprehensively find related articles in several languages, they analyzed the term-relatedness based on 13 million query logs of a search engine that comprised native and transliterated queries. These studies assumed that a single record in a database includes multilingual texts, which, however, cannot be applied to a case where each record is monolingual. Delgado et al. (2018) discussed the problems related to person name disambiguation of web articles written in several languages. They demonstrated that the use of a translator can yield good identification performance for formal texts; however, it works slowly for long texts. Inspired by the above-mentioned related works, we applied a translator to short texts only as a practical solution.

Proposed method

This section describes our novel approach for multilingual author matching across different databases. Figure 1 presents an overview of the proposed method. Suppose that we are provided with two different types of academic databases, namely, source database S and target database T. Our framework assumes that the source database S has a researcher ID assignment system based on manual aggregation of researchers' accomplishments, whereas the target database T has many publication records that are not linked to individuals. The objective of this research is to accurately link each record of T to an existing researcher ID in S.

Notations

A set of researcher IDs corresponding to an English full name x in the source database is represented by S_x. Each researcher ID $\alpha \in S_x$ is accurately associated with a set of records in S, which is denoted as $S(\alpha) \in S$. To obtain this individual's contributions from the other database T, the full name x and its abbreviations with initials can be used as the search queries. For example, suppose that the full name x in S is "Takashi WATANABE," the name variations in T can be as follows: "Takashi Watanabe," "T. Watanabe," and "T. A. Watanabe." A set of records in T, whose author names belong to the variation of x, is represented

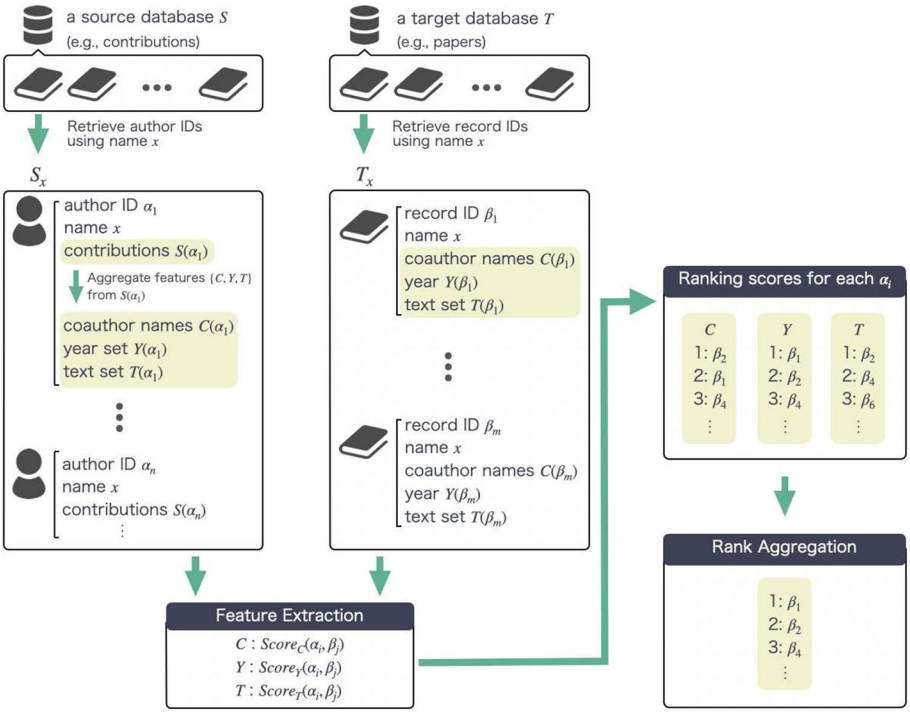


Fig. 1 Overview of the proposed method

by $T_x = \{\beta\}$, in which β represents a single record in T. In the next subsection, we calculate the pairwise similarity between the scholarly contributions $S(\alpha)$ of a researcher ID $\alpha \in S_x$ and a single record $\beta \in T_x$.

Similarity functions

Considering its applicability to any type of scholarly information, our method utilizes the following typical attributes: coauthors, publication dates, and research content words. We derive a similarity measure for each type of these attributes, whose computation is simple for practical use.

Coauthor-based similarity

Coauthors are known as strong features for identifying whether two documents are written by the same individual. Several measures are available for calculating the similarity between two sets. Following the success of conventional studies (Shen et al. 2017), we compare the following three famous measures: Jaccard coefficient, Dice coefficient, and Simpson coefficient. For each $\alpha \in S_x$, let $C(\alpha)$ be a pool of all coauthor names extracted from the contributions in $S(\alpha)$. Similarly, let $C(\beta)$ be a set of coauthor names for $\beta \in T_x$. The similarity between the two sets $C(\alpha)$ and $C(\beta)$ can be calculated as follows:

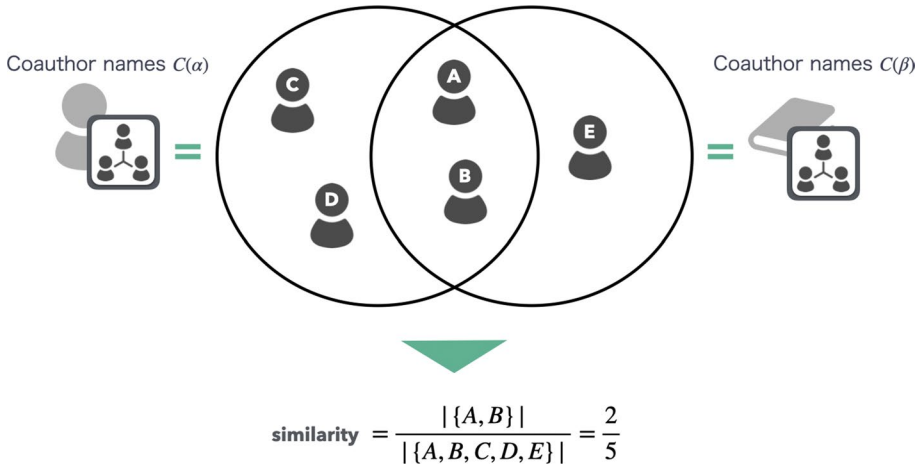


Fig. 2 Example of coauthor-based similarity calculation using the Jaccard coefficient

$$\text{Score}_{\text{co}}^{\text{jaccard}}(\alpha, \beta) = \frac{|C(\alpha) \cap C(\beta)|}{|C(\alpha) \cup C(\beta)|}, \tag{1}$$

$$\text{Score}_{\text{co}}^{\text{dice}}(\alpha, \beta) = \frac{2|C(\alpha) \cap C(\beta)|}{|C(\alpha)| + |C(\beta)|}, \tag{2}$$

$$\text{Score}_{\text{co}}^{\text{simpson}}(\alpha, \beta) = \frac{|C(\alpha) \cap C(\beta)|}{\min(|C(\alpha), C(\beta)|)}. \tag{3}$$

Figure 2 illustrates an example of coauthor-based similarity calculation in the case of using the Jaccard coefficient. In our experiments, we compare the performance of these three measures on author matching.[5]

Publication year-based similarity

If the research activity periods of two authors overlap or are close to each other, the likelihood that the two are the same person is increased. Because a researcher ID $\alpha \in S_x$ is already linked to its scholarly contributions in S, the research activity period of α can be represented using a pool of publication periods. If the publication date of a record $\beta \in T_x$ overlaps with the research activity period of α, the record β could be regarded as an achievement of researcher α's activities. However, no well-known metrics are available to calculate such temporal overlap. We propose two types of measures for calculating the publication year-based similarity.

Let us denote publication years of a record $e \in S(\alpha)$ and a record $\beta \in T_x$ by $Y(e)$ and $Y(\beta)$, respectively. For a given record $\beta \in T_x$ and its author candidates $\alpha \in S_x$, the first

[5] Each coauthor name can have variations in initial abbreviations. We considered two authors whose abbreviated names are the same as the matched entities to calculate the intersection of the two sets.

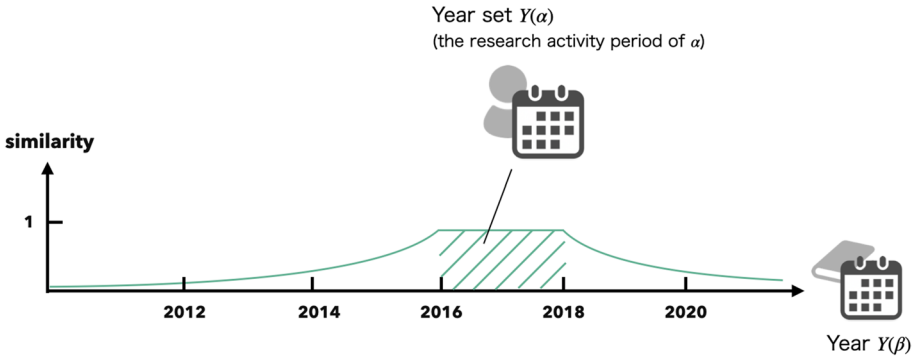


Fig. 3 Outline of calculating the year-based similarity in the case of using Eq. (5), which produces a continuous value

measure outputs a binary value, which indicates whether the record β was published within the activity period of the researcher α or not.

$$\mathrm{Score}^{\mathrm{binary}}_{\mathrm{year}}(\alpha,\beta) = \begin{cases} 1, & \text{if } d(\alpha,\beta) = 0, \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

where

$$d(\alpha,\beta) = \begin{cases} 0, & \text{if } \min\limits_{e\in S(\alpha)} Y(e) \le Y(\beta) \le \max\limits_{e\in S(\alpha)} Y(e), \\ \min\limits_{e\in S(\alpha)} |Y(e) - Y(\beta)|, & \text{otherwise}. \end{cases}$$

The second measure outputs a continuous value, which increases if $Y(\beta)$ is close to the activity periods of the researcher α, as shown in Fig. 3. Specifically, the similarity is calculated as the inverse of the minimum difference between the publication years, as follows:

$$\mathrm{Score}^{\mathrm{continuous}}_{\mathrm{year}}(\alpha,\beta) = \frac{1}{d(\alpha,\beta)+1}. \tag{5}$$

We compare these two measures via experiments.

Content-based similarity

Textual data, such as titles of scholarly contributions, are supposed to reflect the author's research interests. The higher the similarity between the two sets of textual data, the higher the likelihood that the authors are the same individual. Figure 4 shows an overview of the calculation of content-based similarity. To calculate the similarity of texts in a vector space, we present two text vectorizer approaches, which can be easily adopted in multilingual author matching. The first method is to apply a translator[6] to the text in a source database and calculate the TF-IDF vectors using the same feature space for both the target and source databases. Such keyword-based feature extraction can reflect the author's characteristic wording and is

[6] We used Google Cloud Translation API (https://cloud.google.com/translate/).

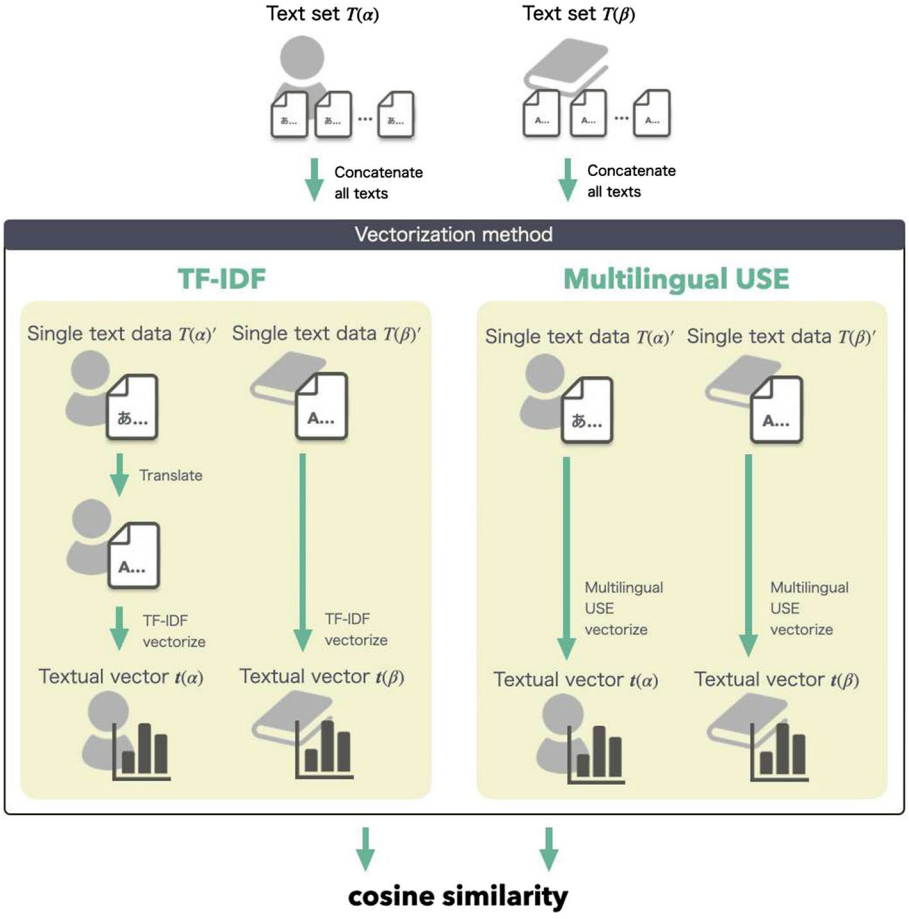


Fig. 4 Two types of methods for vectorizing textual content of records: TF-IDF and Multilingual USE

known to be effective for author identification (Han et al. 2015; Katsurai and Ohmukai 2019). The second method is to use multilingual word embedding, known as Multilingual Universal Sentence Encoder (Multilingual USE) (Yang et al. 2019), which maps the text in several languages in the same vector space. Specifically, the pretrained Multilingual USE model allows 16 languages as input languages, and outputs a 512-dimensional vector for a given text.

For the researcher $\alpha \in S_x$, we first extract all texts (i.e., titles and keywords) from $S(\alpha)$ and concatenate them into a single sentence. The resulting sentence is denoted by $T(\alpha)'$. Similarly, the text of $\beta \in T_x$ is denoted by $T(\beta)'$. Then, by applying either of the two vectorizer methods to $T(\alpha)'$ and $T(\beta)'$, we obtain a single textual vector $\boldsymbol{t}(\alpha)$ and $\boldsymbol{t}(\beta)$. To calculate the content-based similarity between α and β, we use the cosine similarity between the textual vectors, as follows:

$$\mathrm{Score}_{\mathrm{text}}(\alpha, \beta) = \frac{\boldsymbol{t}(\alpha) \cdot \boldsymbol{t}(\beta)}{\|\boldsymbol{t}(\alpha)\| \|\boldsymbol{t}(\beta)\|}, \tag{6}$$

Table 1 Particulars of metadata field used in the experiments

	Source database	Target database	
	KAKEN	DBLP	PubMed
Data types	Project	Article/ inpro- ceedings	MEDLINE article
Language	Japanese	English	English
MetadataMetadata			
Researcher ID	eradCode	–	–
Coauthor names	Member	Author	AuthorList
Publication year	PeriodOfAward	Year	PubDate
Text	Title	Title	ArticleTitle
	Paragraph		Abstract
	Keywords		MeshHeadingList
			ChemicalList

where $\|\cdot\|$ represents the L2 norm a vector. The performance of the two vectorization methods is evaluated through experiments under the following conditions: (i) cosine similarity based on TF-IDF only, (ii) cosine similarity based on Multilingual USE only, and (iii) averaged cosine similarity between TF-IDF and Multilingual USE.

Unsupervised score aggregation

Using each similarity type $k \in \{\text{co, year, text}\}$, for a target researcher $\alpha \in S_x$, we obtain the scores $\{\text{Score}_k(\alpha, \beta)\}_{\beta \in T_x}$, which indicate the likelihood of the occurrence of the same individual's records. To merge the scores of these different types, we utilize an unsupervised score aggregation approach, namely, CombSUM (Fox and Shaw 1994). The effectiveness of CombSUM was demonstrated in our previous study (Katsurai and Ohmukai 2019). We first obtain normalized scores from each similarity measure using Min-Max normalization, as follows:

$$\text{Score}_k(\alpha, \beta) \leftarrow \frac{\text{Score}_k(\alpha, \beta) - \min_k}{\max_k - \min_k}, \quad k \in \{\text{co, year, text}\}, \tag{7}$$

where $\max_k$ and $\min_k$ are the maximum and minimum scores among the set $\{\text{Score}_k(\alpha, \beta); \beta \in T_x\}$, respectively. Then, we calculate the sum of the scores from all similarity measures for each record β as follows:

$$\text{Score}(\alpha, \beta) = \text{Score}_{\text{co}} + \text{Score}_{\text{year}} + \text{Score}_{\text{text}}. \tag{8}$$

Sorting these scores in descending order produces a ranked list of records $T_x = \{\beta\}$. Such a list can reduce the cost for finding the researcher α's work in the target database T.

Datasets

Because our method is developed for practical database management, we evaluated its performance using a large-scale academic database that encompasses all disciplines in Japan, namely, KAKEN, as a source database. To automatically associate English publication

Table 2 Distribution of homography in KAKEN-informatics

# of distinct full names	# of corresponding author IDs	Total
2	3	6
56	2	112
58	–	118

The experiments matched 118 KAKEN author IDs with their DBLP records

records to researchers in KAKEN, we used two well-known public English databases, namely, DBLP and PubMed, as target databases. Table 1 lists the details of metadata (of each database) used in the experiments.

KAKEN dataset

KAKEN is a public database in Japan, which includes project information about research grants provided by the Japan Society for the Promotion of Science, such as the Grants-in-Aid for Scientific Research Program. In KAKEN, each researcher has a unique researcher ID (known as eradCode), and each project is accurately linked to the researcher IDs corresponding to its authors. We considered all 911,724 KAKEN projects registered as of July 2019. Each research project has attributes based on the KAKEN XML definition.[7] It is associated with `id` (project ID), `title`, `field`, `keyword`, `paragraph`, `member`, and `periodOfAward`. The `member` contains `eradCode` (author ID) and `fullName` (author's English name). To construct a testing set of researchers, we extracted the KAKEN projects whose field labels correspond to Informatics or Biology. The total number of researchers who appear in at least one project of Informatics or Biology was 10,383 and 92,339, respectively. We counted the frequency of occurrence of English full names in the pool of researchers and compiled a list of names that correspond to multiple individuals.

To evaluate the performance of multilingual author matching, the ground truth of an individual's English and non-English publication list must be prepared. In Japan, researchers who receive KAKEN grants must submit partial lists of publications published during grant periods. Although a publication list contains the titles of papers, each title is simply written as text and is not linked to the entities of other database records. To automatically find DBLP or PubMed records whose titles match these title strings, we calculated the similarity between two strings, as follows:

$$sim_{title} = 1 - \frac{ldist}{lensum}, \tag{9}$$

where *lensum* is the sum of two title lengths and *ldist* denotes the Levenshtein distance between two title strings. We regarded the records whose string similarities were greater than 0.8 as the same records to construct pseudo ground truth. KAKEN researcher IDs that have at least one record that matches with DBLP or PubMed records were used in our experiments.

[7] https://bitbucket.org/niijp/kaken_definition/src/dac3b303dc90?at=master.

Table 3 Distribution of homography in KAKEN-biology

# of distinct full names	# of corresponding author IDs	Total
1	12	12
1	11	11
4	9	36
8	8	64
10	7	70
22	6	132
46	–	325

The experiments matched 325 KAKEN author IDs with their PubMed records

Because the number of KAKEN researchers in Biology is larger than that in Informatics, we used only the full names that correspond to more than five individuals, in the experimental settings for Biology. The resulting homography researcher set for Informatics and Biology is denoted by *KAKEN-Informatics* and *KAKEN-Biology*, respectively. Tables 2 and 3 show the distributions of the number of researcher IDs for full names in these two KAKEN homography sets. Interestingly, the most popular full name in the field of Biology includes 12 individuals, indicating the problem of author name ambiguity.

DBLP dataset construction

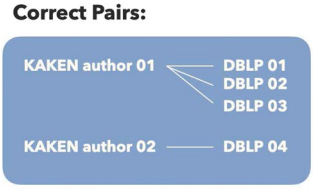
The DBLP computer science bibliography is an English database of open bibliographic information on computer science journals and proceedings (Ley 2009). We collected 4,604,358 DBLP records labeled with article or inproceedings available as of October 2019. Each record contains the metadata of `author` (i.e., coauthor names), `year` (i.e., publication year), and `title`. Very few records in DBLP contain manually inserted author information: most records have no author identification, and each record's author field generally contains only a character string of author names.

DBLP is bound by a rule to store a publication's author names. Specifically, a name is represented in the form of "first name + blank space + last name." If the first name is abbreviated, the initial should always be followed by a period (i.e., dot). Behind a period, there should always be a blank space or a hyphen. In experiments, we compare the author names across the records in the source and target databases according to the above rule. Although DBLP optionally appends a space character and a four-digit number to author names for identifying authors, we ignore this number for simplicity.

Using each full name in the KAKEN-Informatics as a query, we searched for authors who have the same names in DBLP. The total number of ground truth records whose author names appeared in KAKEN-Informatics was 1406, and the average number of records per full name was about 24.

PubMed dataset construction

PubMed provides more than 30 million citations of the MEDLINE database in the fields of biomedicine and life sciences. In the experiments, we used MEDLINE articles, which constitute a large proportion of the whole database. The total number of records was 29,825,494 as of July 2019. Each record has the following attribute types: `AuthorList`



Estimated ranking:

Rank j	KAKEN author 01	Precision@j
Rank 1	DBLP 01 (correct)	1
Rank 2	DBLP 04 (incorrect)	-
Rank 3	DBLP 02 (correct)	0.67
Rank 4	DBLP 03 (correct)	0.75

$$AP = \frac{(1 + 0.67 + 0.75)}{3} \simeq 0.81$$

Fig. 5 Example of AP calculation

(i.e., coauthor names), `PubDate` (i.e., publication year), and textual data seen in `ArticleTitle`, `Abstract`, `MeshHeadingList` and `ChemicalList`. Similar to DBLP, the author field of PubMed records generally contains only a character string of author names and is not identified by any author ID. The total number of ground truth records whose author names appeared in KAKEN-Biology was 4178, and the average number of records per full name was approximately 91.

Experiment

Evaluation measure

As a quantitative evaluation measure, we used mean average precision (MAP), which corresponds to the average of average precision (AP). Figure 5 shows an example of AP calculation in our experiments. This example assumes that when two distinct KAKEN author IDs (01 and 02) are given from a source database as a homography set, their actual DBLP records (01 to 04) are known as ground truth, as shown on the left side of Fig. 5. Our method provides each KAKEN author ID with a ranking of four DBLP records according to the calculation of Eq. (8), as shown on the right side of this figure. DBLP records 01 to 03 should be ranked at the top for KAKEN author 01, whereas DBLP record 04 should be ranked at the top for KAKEN author 02. According to the definition of the precision measure, we obtained the values of 1, 0.67, and 0.75 as precisions for KAKEN author 01. On averaging the precisions, the author ID produces AP, and further averaging the APs over all KAKEN author IDs produces MAP. The larger the value of MAP, the better the performance.

Results

We investigated the influence of different combinations of similarity measures on the MAP using the two experimental settings: matching DBLP records to the researcher IDs of KAKEN-Informatics and matching PubMed records to the researcher IDs of KAKEN-Biology. We name these *KAKEN-DBLP* and *KAKEN-PubMed* settings, respectively. To

Table 4 Results obtained with different combinations of similarity measures in KAKEN-DBLP

Similarity measures	MAP
Coauthor (Jaccard)	0.70
Coauthor (Dice)	0.70
Coauthor (Simpson)	0.70
Publication year (continuous)	0.67
Publication year (binary)	0.65
Content (TF-IDF, titles only)	0.78
Content (TF-IDF, all text)	0.81
Content (Multilingual USE, titles only)	0.78
Content (Multilingual USE, all text)	0.83
Coauthor + year	0.79
Year + content (TF-IDF, titles only)	0.85
Year + content (Multilingual USE, titles only)	0.85
Coauthor + content (TF-IDF, titles only)	0.82
Coauthor + content (Multilingual USE, titles only)	0.82
Coauthor + year + content (TF-IDF, titles only)	0.87
Coauthor + year + content (Multilingual USE, titles only)	0.88
Coauthor + year + content (average of TF-IDF and Multilingual USE, titles only)	0.89
Coauthor + content (average of TF-IDF and Multilingual USE, all text)	0.89
Coauthor + year + content (average of TF-IDF and Multilingual USE, all text)	***0.92***

Table 5 Results obtained with different combinations of similarity measures in KAKEN-PubMed

Similarity type	MAP
Coauthor (Jaccard)	0.68
Coauthor (Dice)	0.68
Coauthor (Simpson)	0.68
Publication year (continuous)	0.22
Publication year (binary)	0.22
Content (TF-IDF titles only)	0.64
Content (TF-IDF, all text)	0.88
Content (Multilingual USE, titles only)	0.61
Content (Multilingual USE, all text)	0.74
Coauthor + year	0.65
Year + content (TF-IDF, titles only)	0.67
Year + content (Multilingual USE, titles only)	0.65
Coauthor + content (TF-IDF, titles only)	0.82
Coauthor + content (Multilingual USE, titles only)	0.77
Coauthor + year + content (TF-IDF, titles only)	0.80
Coauthor + year + content (Multilingual USE, titles only)	0.77
Coauthor + year + content (average of TF-IDF and Multilingual USE, titles only)	0.82
Coauthor + content (average of TF-IDF and Multilingual USE, all text)	***0.93***
Coauthor + year + content (average of TF-IDF and Multilingual USE, all text)	0.87

comprehensively investigate the effectiveness of the content-based similarity measures, we conducted experiments for the following cases. (a)"titles only": we used only titles of records in both datasets as textual data; and (b) "all text": we used all text that is available in both datasets. DBLP records contain no keyword, whereas PubMed records are associated with rich textual data as shown in Table 1.

Tables 4 and 5 summarize the results of possible combinations of similarity measures in KAKEN-DBLP and KAKEN-PubMed settings, respectively, in which the best scores are highlighted in bold. Focusing on the performance of a single feature type, coauthor-based similarity measures exhibited stable performance for both settings, and we observed that the type of coefficients (i.e., Jaccard, Dice, or Simpson) does not affect the performance. On the contrary, the publication year-based similarity measure worked differently in the two settings: its performance in KAKEN-DBLP was almost the same as that of the coauthor-based similarity measure but was significantly degraded in KAKEN-PubMed compared with that of other similarity measures. This is because the KAKEN-PubMed setting contained many negative samples due to the large size of each fullname's homography, in which the scalar-based measure (i.e., year only) especially cannot rank similarities of publication year well. Thereafter, for integrating different types of similarity measures, we chose the Jaccard coefficient for coauthor-based similarity, a continuous index for publication year-based similarity.

In the case of applying TF-IDF or Multilingual USE to titles, it is evident that there is no significant difference between them. The use all text available in the datasets delivered better performance than the use of "titles only." TF-IDF with all text demonstrated a significantly large MAP in the KAKEN-PubMed setting. We can consider that TF-IDF performs well in discriminating individuals when their records are associated with many keywords.

The integration of similarity measures often delivered better performance than a single measure, implying that the shortcomings of a single similarity were effectively compensated for by other measures. Although the publication year-based similarity did not perform effectively by itself, it became an additional useful feature for other measures in the KAKEN-DBLP setting. In contrast, integrating the year-based similarity degraded the overall performance in the KAKEN-PubMed setting. Although our current method sums three similarity scores using equal weights (see Eq. 8), tuning the weight can be an effective solution to improve the overall performance. However, it is also difficult to determine weights that are valid to all dataset combinations. Thus, we can conclude that integrating all similarity measures is currently the most reasonable approach available for a certain pair of datasets, and how to weight each similarity measure should be investigated in our future works. Furthermore, we should develop an approach that uses the publication year information to remove candidates whose research periods are clearly different over a span of several decades.

Results for specific full names

Table 6 shows the APs of seven full names, obtained using "coauthor + year + content (average, titles only)," in which (*a*)–(*d*) and (*e*)–(*g*) are the results of KAKEN-DBLP and KAKEN-PubMed settings, respectively. The proposed method performed well for the homography of full names (*a*), (*b*), (*e*), and (*f*). We obtained large APs when the fields of source KAKEN homography researchers are diverse, and when each researcher has a large number of records in a target database. Because the MAP evaluation is affected by the number of positive examples, the performance of our method also degrades when the target

Table 6 Results for seven full names, obtained using "coauthor + year + content (average)"

KAKEN full name	Researcher ID	AP	Research field	# of records
(*a*) Hashimoto hideki	a_1	1	Human interface	4
	a_2	1	Big data analysis	2
	a_3	0.93	Optimization	11
(*b*) Numao masayuki	b_1	0.99	Data mining of biological signals	23
	b_2	0.70	Data mining in medical welfare	7
(*c*) Murata masaki	c_1	1	Natural language processing	8
	c_2	0.17	Natural language processing	1
(*d*) Shibata naoki	d_1	1	Sensor network	33
	d_2	0.17	Citation network	1
(*e*) Tanaka hiroyuki	e_1	1	Fisheries chemistry	6
	e_2	1	Chemical pharmacy	3
	e_3	1	Plant molecular biology	1
	e_4	0.94	Regenerative medicine	14
	e_5	0.93	Pathology	12
	e_6	0.92	Bioorganic chemistry	11
	e_7	0.87	Structural biochemistry	10
	e_8	0.85	Breeding science	6
	e_9	0.82	Biological pharmacy	10
	e_{10}	0.82	Cardiovascular surgery	5
	e_{11}	0.75	Pediatrics	8
	e_{12}	0.58	Physical anthropology	2
(*f*) Saito akira	f_1	1	Periodontal dentistry	6
	f_2	1	Sports science	4
	f_3	1	Plastic surgery	3
	f_4	1	Chemical pharmacy	1
	f_5	1	Applied health science	1
	f_6	0.93	Respiratory medicine	17
(*g*) Tanaka hirokazu	g_1	0.95	Hematology	27
	g_2	0.91	Obstetrics and gynecology	8
	g_3	0.83	Neurophysiology	5
	g_4	0.71	Cardiovascular surgery	3
	g_5	0.24	Plant molecular biology	3
	g_6	0.22	Immunology	2
	g_7	0.09	Obstetrics and gynecology	1
	g_8	0.04	Liver surgery	1

The number of records represents the corresponding individual's actual publications in a target database (DBLP or PubMed)

researcher has only a few records. On the contrary, the results of (*c*), (*d*) and (*g*) had small APs due to the overlap of research fields and because only a few records are written by one of the researchers. For example, in (g), the fields Hematology and Immunology are relevant to each other and tend to share numerous technical terms; the same is true for the fields of Cardiovascular surgery and Liver surgery as well. Our current content-based approach

exhibits difficulties in capturing such subtle differences in technical meanings from short texts. This limitation can be potentially overcome using a multilingual text encoder that particularly learns scientific terms. In addition, profiling each KAKEN researcher in more detail would be effective. These issues will be addressed in our future works.

Conclusions

An approach for multilingual author matching across different academic databases was developed in this study. Considering the applicability to actual database management, our method is unsupervised, and we use simple similarity measures that can be easily implemented. For calculating the textual relevance between English and non-English records, we presented both a translator-based TF-IDF approach and a multilingual sentence embedding approach. In the experiments on KAKEN, DBLP, and PubMed, a similarity measure that integrates all types of similarity (i.e., coauthor-based, publication year-based, and content-based) achieved stable performance in both KAKEN-DBLP and KAKEN-PubMed settings and is currently the most reasonable approach. The translated TF-IDF and sentence embeddings work differently with each dataset, and we found that obtaining their average provides a stable index in any domain. In addition, the use of "all texts" available in each database's records delivered better performance than the use of "titles only."

Although our study presented a base for multilingual author matching, it has further scope for improvement. First, when a single measure was used in our experiments, the performance of publication year-based similarity was lower than that of other similarity measures. Therefore, an effective index must be developed that can measure the temporal overlaps between the records, or an approach that can filter candidates must be created. Furthermore, the matching performance degraded when a target set of KAKEN homography included individuals whose research fields were similar. This is possibly due to the inability to capture the subtle differences in technical nuances from short texts. This problem can be solved using a multilingual text encoder that specializes in learning technical terms. The future scope of our study includes effective re-training of a sentence embedding model on a large-scale scientific corpus. Furthermore, we plan to develop more sophisticated techniques for integrating different similarity measures and profiling each author's research field in detail.

Acknowledgements This research was partly supported by JSPS KAKENHI Grant Number 20H04484, JST ACT-I grant number JPMJPR18UC, JST ACT-X grant number JPMJAX1909, and ROIS NII Open Collaborative Research 2020-20S0405.

References

Amano, T., González-Varo, J. P., & Sutherland, W. J. (2016). Languages are still a major barrier to global science. *PLoS Biology*, *14*(12), e2000933.

Bhattacharya, I., & Getoor, L. (2007). Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *1*(1), 5-es.

Cota, R. G., Ferreira, A. A., Nascimento, C., Gonçalves, M. A., & Laender, A. H. (2010). An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations. *Journal of the American Society for Information Science and Technology*, *61*(9), 1853–1870.

Delgado, A. D., Martínez, R., Montalvo, S., & Fresno, V. (2018). Person name disambiguation on the web in a multilingual context. *Information Sciences*, *465*, 373–387.

Ferreira, A. A., Gonçalves, M. A., & Laender, A. H. (2012). A brief survey of automatic methods for author name disambiguation. *ACM SIGMOD Record*, *41*(2), 15–26.

Fox, E. A., & Shaw, J. A. (1994). Combination of multiple searches. In *TREC-2*, pp. 243–252.

Gupta, P., Bali, K., Banchs, R. E., Choudhury, M., & Rosso, P. (2014). Query expansion for mixed-script information retrieval. In *Proceedings of the 37th international ACM SIGIR conference on research and development in information retrieval* (pp. 677–686), ACM.

Han, D., Liu, S., Hu, Y., Wang, B., & Sun, Y. (2015). ELM-based name disambiguation in bibliography. *World Wide Web*, *18*(2), 253–263.

Han, H., Zha, H., & Giles, C. L. (2005). Name disambiguation in author citations using a k-way spectral clustering method. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'05)* (pp. 334–343), IEEE.

Jung, J. J. (2013). Cross-lingual query expansion in multilingual folksonomies: A case study on Flickr. *Knowledge-Based Systems*, *42*, 60–67.

Katsurai, M., & Ohmukai, I. (2019). Author matching across different academic databases: Aggregating simple feature-based rankings. In *2019 ACM/IEEE joint conference on digital libraries (JCDL)* (pp. 279–282).

Ley, M. (2009). DBLP: Some lessons learned. *Proceedings of the VLDB Endowment*, *2*(2), 1493–1500.

Long, N. H., & Jung, J. J. (2015). Privacy-aware framework for matching online social identities in multiple social networking services. *Cybernetics and Systems*, *46*(1–2), 69–83. https://doi.org/10.1080/01969722.2015.1007737.

Meneghini, R., & Packer, A. L. (2007). Is there science beyond English? *EMBO Reports*, *8*(2), 112–116.

Müller, M.-C., Reitz, F., & Roy, N. (2017). Data sets for author name disambiguation: An empirical analysis and a new resource. *Scientometrics*, *111*(3), 1467–1500.

Salager-Meyer, F. (2014). Writing and publishing in peripheral scholarly journals: How to enhance the global influence of multilingual scholars? *Journal of English for Academic Purposes*, *13*, 78–82. https://doi.org/10.1016/j.jeap.2013.11.003.

Shen, Q., Wu, T., Yang, H., Wu, Y., Qu, H., & Cui, W. (2017). NameClarifier: A visual analytics system for author name disambiguation. *IEEE Transactions on Visualization and Computer Graphics*, *1*, 141–150.

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Abrego, G. H., Yuan, S., Tar, C., Sung, Y.-H., et al. (2019). Multilingual universal sentence encoder for semantic retrieval. arXiv preprint arXiv:190704307.