RESEARCH ARTICLE

Identifying constitutive articles of cumulative dissertation theses by bilingual text similarity. Evaluation of similarity methods on a new short text task

Paul Donner

German Centre for Higher Education Research and Science Studies (DZHW), Dept. 2 'Research System and Science Dynamics', Schützenstr. 6a, 10117 Berlin, Germany

Keywords: bilingual text similarity, cross-language information retrieval, cumulative dissertation, doctoral thesis

ABSTRACT

Cumulative dissertations are doctoral theses comprised of multiple published articles. For studies of publication activity and citation impact of early career researchers, it is important to identify these articles and link them to their associated theses. Using a new benchmark data set, this paper reports on experiments of measuring the bilingual textual similarity between, on the one hand, titles and keywords of doctoral theses, and, on the other hand, articles' titles and abstracts. The tested methods are cosine similarity and L_1 distance in the Vector Space Model (VSM) as baselines, the language-indifferent methods Latent Semantic Analysis (LSA) and trigram similarity, and the language-aware methods fastText and Random Indexing (RI). LSA and RI, two supervised methods, were trained on a purposively collected bilingual scientific parallel text corpus. The results show that the VSM baselines and the RI method perform best but that the VSM method is unsuitable for cross-language similarity due to its inherent monolingual bias.

Citation: Donner, P. (2021). Identifying constitutive articles of cumulative dissertation theses by bilingual text similarity. Evaluation of similarity methods on a new short text task. *Quantitative Science Studies*, *2*(3), 1071–1091. https://doi.org/10.1162/qss_a_00152

DOI:
https://doi.org/10.1162/qss_a_00152

Supporting Information:
https://doi.org/10.1162/qss_a_00152

Received: 18 May 2021
Accepted: 11 August 2021

Corresponding Author:
Paul Donner
donner@dzhw.eu

Handling Editor:
Vincent Larivière

1. INTRODUCTION

1.1. Background and Motivation

What is the contribution of early career researchers (ECRs) to a country's research output? This question is currently of high science-political interest in Germany and of similarly high practical difficulty to answer (Consortium for the National Report on Junior Scholars, 2017, p. 19). The training of qualified research workers is widely regarded as a core mission of universities and accordingly the performance of universities and departments with respect to the training of ECRs plays a prominent role in research evaluation systems. Yet, despite the acknowledged interest in performance of ECRs in terms of scientific output—publications and their citations—this facet of performance, research output, has so far not become part of university evaluation systems or national scale monitoring instruments. Beyond these science-political considerations, the research contribution and performance of ECRs is intrinsically interesting. Comprehensive performance data would enable longitudinal observation and trend detection,

comparisons between ECRs of different disciplines, and perhaps the detection of effects of political interventions or different institutional conditions across legislatures (federal states) and organizations (universities) pertaining to ECR performance.

PhD theses are the primary published research output of a completed PhD degree in Germany because the full thesis needs to be published in some format for a degree to be conferred. The publication format may be a regular book with a scientific publishing house or a digital document deposited at a university repository. The regulations vary and are locally determined at the university or department level. Many doctoral students also publish in periodicals and contribute to conference proceedings and edited book chapters. These articles by one PhD candidate might be collated, supplemented with introduction and conclusion material, and submitted as a cumulative PhD thesis, which still needs to be published as a unit. The other class of theses is monograph theses, which have been designed as one single work from the outset and which are not published in parts otherwise.

The importance of cumulative (i.e., article-based) dissertations in Germany can be seen from a number of recent surveys among PhD students and graduates. A 2014 survey of PhD graduates found that 14.5% completed a cumulative thesis while 84% handed in a monograph thesis (Brandt, Briedis et al., 2020). This survey also inquired about the number of cumulative articles. The mean number was 4.0 and the median 3. An analysis by the German Federal Statistical Office found that in 2015 the share of PhD students working on a cumulative dissertation was 23% while 77% worked on monograph theses (Hähnel & Schmiedel, 2016). For seven science domains, the figures for cumulative dissertations varied between 13% in language and cultural studies and 60% in agricultural and food sciences. A 2018 survey asked PhD students on the planned format of their thesis: 25% planned a cumulative thesis, 57% a monograph thesis, and the bulk of the remainder were undecided (Adrian, Ambrasat et al., 2020). In summary, while the monograph dissertation remains the more common format, the importance of cumulative dissertations is substantial and increasing. Therefore, both theses and constitutive articles need to be taken into account in studies of knowledge production and the citation impact of ECRs.

Our study aims to partially assemble the technical requirements for bibliometric studies of the output of doctoral degree holders, which so far are lacking. In particular, we evaluate several methods of short text similarity calculation on their ability to support the identification of the elemental articles of cumulative PhD theses. This is only one component of a complete PhD candidate article identification system, as will be discussed below, but a centrally important one. It is necessary to identify thesis-related articles in the first place because, in the case of Germany, there is no public register of PhD students or graduates, nor is there a comprehensive source of persistent identifiers of PhD students or graduates and their nonthesis articles[1]. With a central register containing PhD candidate names and their university affiliations, it would be possible to do comprehensive and targeted searches of publication databases. Yet, without due caution the results would probably contain inaccuracies. As different persons can have the same name and a single person can publish with different names, an author name disambiguation system is a general requirement for high-quality author-level data. Not all publication databases have such systems, and some vendors do not report on their matching quality. Even if there were a perfect author name disambiguation

[1] Germany does not maintain a central register of active PhD students or graduates (Fräßdorf & Fräßdorf, 2016) and universities have only been required to systematically and comprehensively collect data on current doctoral students since 2017 (Brauer, Oelsner, & Boelter, 2019). These new data are decentralized, not public, and only cover the period since 2017.

system, one would need a reliable automatic system for identifying the correct author record among candidates (identity matching) because of name homonymy (different persons with the same name)[2]. Another reason why information on names and university affiliations is insufficient for finding thesis-associated articles is that far from all doctoral candidates are formally affiliated with the universities at which they obtain their degree (e.g., Gerhardt, Briede, & Mues, 2005).

It is possible to bypass the author identity problem by simply considering the author names of a thesis and of candidate-associated articles as only one feature of a larger, jointly used feature set. In other words, instead of matching articles via their disambiguated authors, one can match theses and articles via author names (rather than author identities) and a suite of other features, such as information on affiliation, publication year, and topic. This is the approach pursued in the project of which this study is a part[3]. We anticipate handling the matching of thesis records and candidate article records by supervised classification algorithms. As a recent example of this approach, Heinisch, Koenig, and Otto (2020), in a thematically closely related study, perform machine learning-based record linkage on bibliographic data on German doctorate recipients with administrative labor market data to trace their career outcomes.

An important feature for matching cumulative theses and their constitutive articles is their topical similarity (Echeverria, Stuart, & Blanke, 2015; Zamudio Igami, Bressiani, & Mugnaini, 2014) and we investigate in this paper the optimal computation of topical similarity under the specific conditions of the task at hand. In other words, due to the complexity of the subtask of finding a good topic similarity measure for this specific application, in this contribution we cannot address the entire PhD candidate article identification system, but focus on this important subtask. It is important to mention this context to preclude the misapprehension that the similarity measures studied will be used in isolation to identify articles constituting cumulative PhD theses. The basic premise is that articles on the same or a similar topic are more likely to be proper parts of a given thesis than topically remote or unrelated articles, even by the same authors. In addition, topical similarity can be useful to distinguish between articles by different authors with the same name or name abbreviation in the aforementioned automatic classification stage. The results obtained in this study could therefore also inform future research in author disambiguation. Topical similarity is most conveniently operationalized as textual similarity. While other operationalizations are also appropriate, such as distance in the citation network, the basic data for such approaches is not directly available in the dissertation bibliographic data at hand.

[2] To give an extreme example, between 1996 and 2016, according to data collected from the German National Library, which was deduplicated and excludes medical theses, there are 48 persons named Thomas Müller who have authored a doctoral thesis. Of these, two different ones graduated from Heidelberg University in 1999.

[3] Another approach would be to start with the available full-text electronic documents, apply automatic reference identification procedures to extract the cited sources and use these to identify associated articles by the thesis authors. This seems a promising alternative, albeit with the one drawback that some articles that have not been published at the time of the handing in of the dissertation are typically only cited in a provisional way. The larger problem is external to the data itself. As it stands, by no means all dissertations are published as publicly accessible electronic full-text versions. A reference extraction and matching approach would hence either be limited in coverage or need to collect and prepare theses published by publishing houses in book format or deposit copies from libraries. The effort required for this alternative approach was prohibitive in our project, so we decided to work with bibliographic data only.

1.2. Contribution of This Study

In this study we are concerned with methods of measuring topical similarity based on text in the domain of scientific research. Due to the nature of the specific task and the available data, several important considerations need to be addressed.

1. The best available dissertation data (national library bibliographic records) only contain the titles and, for a subset, content descriptor terms assigned by catalogers. Dissertation titles can be quite succinct—two examples from the data set introduced later are "On operads" and "Fairer Austausch" ("Fair exchange"). This is less of a problem for the candidate associated articles to be matched because their metadata usually also contains an abstract. Therefore this task is an example of the short text similarity problem, which is an area of intense specialized research at the intersection of information retrieval and natural language processing in recent years (Kenter & De Rijke, 2015). The difficulty in calculating similarities for short texts is that two short texts on the same topic are not very likely to use the same terms (the vocabulary mismatch problem). Methods based on exact lexical matching of terms are thus likely to be inaccurate because of the restricted amount of information.
2. The textual data are domain specific, as they all are formal reports of scientific research. Scientific text usually contains many specialized technical terms and may use some words with specific meanings other than those in common language use. Therefore, methods and language resources designed for domain-general text, such as news, might not be ideal.
3. Dissertation theses from Germany are typically written in either German or English, other languages being uncommon. For cumulative theses, the incorporated articles need not be in the same language as the thesis title might indicate. It follows that a text similarity method should be able to measure similarity across different languages.

The present paper reports on a study in applied cross-language information retrieval (CLIR) for the purpose of science studies. No new methods are developed, but existing methods are applied to a novel task. The above presented combination of specific factors of the nature of the task means that we cannot simply rely on prior descriptions of the performance of text similarity methods, as these were evaluated on very different problems. Measuring the textual similarity between doctoral theses and their possible constitutive articles on the basis of bibliographic data in a cross-language setting has to our knowledge not been studied before. It was therefore necessary to collect an appropriate ground truth sample to evaluate the studied methods. Furthermore, we also collected domain-specific translated texts to train the tested supervised methods on appropriate data. As the conventional evaluation metrics cannot be applied because of the cases for which no matches should be retrieved (monograph theses—theses without constitutive articles), the choice of appropriate metrics is discussed in some detail.

While this study is concerned with measuring textual similarity between doctoral theses and associated articles, the task of semantic similarity calculation between short representations of scientific texts is of wide applicability in science studies. The calculation of text similarity in bilingual scenarios is of particular importance to those national science systems where English is not the native language and where much research is published in other languages, for which it is crucial to determine links with the international English research literature. It should therefore be noted that even though we only consider the specific scenario of German and English language publications, the methods studied here can be used for any language combination.

The paper proceeds as follows. In the next section we review the related literature. In Section 3 we describe the data sets used in this study. Data preprocessing and the various tested text similarity methods are treated in Section 4, followed by the presentation of our results (Section 5) and a discussion of these findings (Section 6).

2. LITERATURE REVIEW

We focus here, first, on prior research in the paradigm of distributional semantics[4] for CLIR. We make this restriction because of the decisive advantage of these methods, which is that they do not require lexical matches to calculate text similarities. This is crucial for the task of short text similarity calculation, where the probability that two compared texts include the same terms is inherently small, quite independent of their true topical similarity. Second, we also review the application of the selected methods for similarity calculation in the field of scientometrics in general (beyond cross-language information retrieval) to provide a more specific context for the use of these methods in the present study.

This line of research was inaugurated with the Latent Semantic Analysis (LSA) model (Deerwester, Dumais et al., 1990) which was extended for cross-language retrieval by Dumais, Letsche et al. (1997). LSA applies statistical dimension reduction (singular value decomposition, SVD) to the sparse weighted term-document matrix created from a text corpus to obtain a smaller and dense "semantic" vector space representation in which both terms and documents are located. For all input documents and terms, profiles of factor weights over latent extracted factors are obtained that characterize these entities based on observed term co-occurrences in the data. LSA has found significant use in the field of scientometrics. Landauer, Laham, and Derr (2004) illustrate the use of LSA for visualizing large numbers of scientific documents by applying the method to six annual volumes of full texts of papers from PNAS. This study highlights the possibilities of interactive, user-adjustable displays of documents. The study by Natale, Fiore, and Hofherr (2012) studied scientific knowledge production on aquaculture using LSA and other quantitative publication analysis methods. This is an interesting case study because LSA as a topic identification method was triangulated with topic modeling and cocitation analysis on the same corpus of documents. Article titles and keywords were used as inputs for LSA and the similarity values between words from the semantic space were visualized with multidimensional scaling. Important for scientometric applications is that the LSA method is not restricted to textual term data, which was exploited by Mitesser, Heinz et al. (2008) and Mitesser (2008), who applied SVD to matrices encoding papers and cited references of volumes of journals to measure the topical diversity of research and its temporal development, assuming topical structure to be implicit in the patterns of cited literature.

Random Indexing (RI) is a direct alternative to LSA with lower computational demands, meaning that it can be applied to much larger corpora. Sahlgren and Karlgren (2005) used the RI approach to CLIR for the task of automatic dictionary acquisition from parallel (translated) texts in two languages. Moen and Marsi (2013) experimentally studied the performance of RI in *ad hoc* monolingual and cross-language retrieval (German queries, English documents) on standard evaluation data sets. Their variant of the RI method only used a translation dictionary but no aligned translated texts. RI compared unfavorably to the standard VSM and dictionary-based query translation in CLIR[5]. To this day, there is little empirical research on

[4] Or word embeddings, or wordspace, or continuous word vectors, etc. The terminology has not yet stabilized.

[5] Contrary to the standard RI method (Sahlgren, 2005), the authors started out with assigning fixed index vectors to terms, using the same vector for the terms in both languages, rather than starting with index vectors for documents and constructing term index vectors from the document vectors.

CLIR applications of RI and none on short text similarity, to the best of our knowledge. RI has been introduced into the domain of science studies by Wang and Koopman (2017), describing the application of the method to a benchmarking data set used for testing different approaches to scientific document clustering for automatic data-driven research topic detection. A critical particuliarity of their method is that each document is represented by features of distinct types, namely topic terms and phrases extracted from titles and abstracts, author names, journal ISSNs, keywords, and citations. For each such entity, a 600-dimension real-valued vector representation is learned by random projection from their co-occurrences in the corpus. Next, a vector representation for every document is calculated as the weighted centroid of the vectors of its constituting entities. Clustering algorithms are then applied, one to the set of semantic document vectors directly, and another one to a network of similarity values of each document to its nearest neighbors, in which similarities are calculated as the cosine of the document vectors. Their implementation of RI is further developed in Koopman, Wang, and Englebienne (2019) with improved entity weighting and entity vector projection giving better representations. This study showed the application of the method to a different task of relevance to scientometrics: automatic labeling of documents with terms from a large controlled vocabulary. A version of RI was benchmarked against competing state-of-the-art word embedding methods, trained on the same data, at predicting known withheld Medical Subject Headings for biomedical papers, in which it achieved good results.

Vulić and Moens (2015) introduced a comprehensive CLIR system using word embeddings of several languages in one common vector space, which they call "shared inter-lingual embedding space." They point out that in the word embedding retrieval paradigm, monolingual search and cross-lingual search can be integrated into one system in which search within a single language only uses that part of the system relating to one language. Cross-lingual and monolingual search in any of the supported languages are combined seamlessly in multilingual word embedding-based systems, obviating the need for query expansion or result list merging inherent in machine translation-based systems. The study also demonstrated that bilingual embeddings viable for cross-language *ad hoc* retrieval can be obtained from document-aligned parallel corpora and that finer-grained information, such as sentence or word alignments, is not required. The system of Vulić and Moens (2015) relies on bilingual pseudodocuments, documents formed by merging and shuffling terms from aligned documents in two languages, similar to the method of cross-language LSA (Dumais et al., 1997) and showed very competitive results on standard benchmarking data sets for *ad hoc* CLIR, outperforming prior state of the art models in their test setting (English and Dutch queries and documents). Their results further show that for constructing the document-level representations from the term vectors in the word embeddings CLIR paradigm, a standard term weighting approach outperforms an unweighted additive approach to composition.

To conclude this overview, interested readers are directed to Ruder, Vulić, and Søgaard (2019) for a comprehensive survey of cross-language word embedding models, categorizing them by the required training data according to type of alignment (word, sentence, document) and the comparability (translated or comparable).

3. DATA

3.1. Bilingual Dissertation-Article Pairs

Because the objective of this study is to compare methods of measuring text similarity between cumulative doctoral theses and the articles they are comprised of, we created a manually curated ground truth data set of doctoral theses accepted at German universities and their

associated Scopus-covered articles which we use to evaluate the performance of the methods. We chose the bibliographical and bibliometric database Scopus as a source for article-level bibliographic data, as many thesis-related articles are published in the German language and Scopus covers more German-language literature compared to Web of Science. For the period from 1996 to 2017, Scopus contained around 694,000 German-language article records and Web of Science around 500,000 records.

The German National Library (Deutsche Nationalbibliothek, DNB) catalog currently provides the most comprehensive source of data on German dissertation theses. As part of its legal mandate the DNB collects and catalogs all works published in Germany and universities regularly submit dissertation and habilitation theses as deposit copies to the DNB. The DNB collection mandate extends to theses accepted at German universities but published by foreign publishing houses. Nevertheless, there is no reliable information on the completeness of the DNB thesis collection. The DNB catalog clearly identifies all dissertation theses but may contain multiple versions of one thesis, such as a print version, a digital version, and a commercially published version. These versions of one work are not always linked and need to be deduplicated for analytical purposes. DNB dissertation data has recently been used several times in scientometric research (Heinisch & Buenstorf, 2018; Heinisch et al., 2020)[6]. DNB dissertation data are a viable, if challenging, data source for studies on German PhD theses and there is at present no better large-scale source for German PhD thesis data.

Basic bibliographic data for dissertation theses was therefore obtained from DNB catalog records. Records for all PhD dissertations from the German National Library online catalog were obtained in April 2019 using a search restriction in the university publications field of "diss*", as recommended by the catalog usage instructions, and publication year range 1996 to 2018. Records were downloaded by subject fields in the CSV format option, except for the subject medicine[7]. In this first step 534,925 records were obtained. In a second step, the author name and work title field were cleaned and the university information extracted and normalized and non-German university records excluded. We also excluded records assigned to medicine as a first subject class, which were downloaded because they were assigned to other classes as well. As the data set often contained more than one version of a particular thesis because different formats and editions were cataloged, these were carefully deduplicated. In this process, as far as possible the records containing the most complete data and describing the temporally earliest version were retained as the primary records. Variant records were also retained separately. This reduced the data set to 361,655 records, only a small part of which is used in this study.

After these cleaning operations, the DNB dissertation data set contains the bibliographic information for probably nearly all German nonmedical PhD theses in the period covered. To construct the ground truth data set, the next steps were to identify cumulative theses in the processed DNB data, to extract the bibliographic information of constitutive articles of the cumulative theses, and to link these article-level records with Scopus records.

[6] For a very limited subset of recent cumulative theses, the DNB data contains information on included articles if these previously published works are completely incorporated in unchanged form. If the full-texts of the theses and the candidate articles are available, the true positive articles should be almost strict subsegments of the thesis they are part of. This suggests that pairs of cumulative dissertation and constituent articles could be ideal true positive gold standards for plagiarism detection methods.

[7] German "Dr. med." degree dissertations are considered incommensurable to other doctoral degree theses (Senatskommission für Klinische Forschung, Deutsche Forschungsgemeinschaft, 2010; Wissenschaftsrat, 2014).

Table 1. Dissertation thesis title language and article language for data set of cumulative dissertations

	Article language	
Thesis language	de	en
de	161	586
en	7	695

For the first of the above steps, the identification of a sample of cumulative dissertations, we proceeded as follows. In general, the DNB records do not indicate the type (cumulative or monographical) of dissertations. However, we found and used a small number of DNB records from the above data set containing the phrase "kumulative Dissertation" in the title of the record. These were mostly from a single university. Our second approach was to use the full-text URLs in the DNB data. For those DNB records that contained such a URL we attempted to download the full-text PDF file, if successful, extracted the plain text, and indexed it for searching. While we were able to download many full-texts, the majority of university repository URLs turned out to be outdated and unreachable. We were able to obtain 36,640 thesis full-texts, which were searched for keywords and phrases indicating a cumulative thesis.

As a third approach, we randomly sampled universities and searched their online publication repositories for dissertations containing keywords or phrases indicating cumulative theses. For promising-looking hits, the thesis full-texts were downloaded for examination.

For all theses identified as possible cumulative dissertations through these methods we obtained the published full-texts via university repositories. We manually searched all downloaded full-text PDFs for explicit statements about articles associated with a cumulative thesis. For articles that are described by the thesis authors as being part of the thesis or that appear as chapters, we extracted the corresponding bibliographic data. We thus used three independent and complementary methods to identify theses which might contain information on constitutive articles.

Next, we manually searched for the identified associated articles from the cumulative theses in a snapshot of the Scopus bibliometric database from spring 2019 and assigned the Scopus item identifier to the extracted article records. Only Scopus items with the document types article, review, conference paper, chapter, or book were retained. We also kept track of all examined theses for which no associated articles were indicated in the full-texts. These are also included in the ground truth data as negative cases. This sample is therefore only a convenience sample and not a statistically representative random sample of a population[8].

The resulting ground truth data set contains 1,181 doctoral thesis records, of which 771 refer to theses with German titles and 410 to theses with English titles. All thesis records are described by bibliographic information from the DNB. Of these records, 449 were identified as cumulative doctoral theses, but 21 of them did not have any Scopus-covered articles. Of the 428 cumulative theses with Scopus-indexed constituent articles, 218 had German titles and 210 had English titles. A total of 732 theses were identified as standalone theses without any incorporated articles. There were 1,499 pairs of theses and Scopus-contained articles out of 1,946 thesis-article pairs in total. The Scopus coverage of this data set's thesis-associated articles is approximately 77%. The cross-tabulation of thesis language and article language for the subset of the final data set with Scopus-indexed articles is shown in Table 1. Note that throughout the remainder of the paper we abbreviate German and English in the tables as

[8] The data set is available at Donner (2021b).

"de" and "en," respectively. It can be seen that among German-language theses there is a preponderance of English-language articles and while most articles of cumulative theses with English titles were also written in English, there are also a few German articles among them. However, it should be kept in mind that in the Scopus data there are always English titles and abstracts, while German titles are present additionally for German-language articles. Thus the cross-language problem is possibly at least partly mitigated by the presence of both German and English text for German articles in Scopus.

This test data set consists, for the doctoral theses, of author names, thesis title in either German or English, German language keywords assigned by DNB catalogers (only partial coverage), publication year, and university. Article bibliographic data from Scopus are comprised of author names, title in English (always present) and German (sometimes present), English abstract, publication year, and disambiguated German institution (if present). For the text similarity task of this paper, only the thesis title and keywords and the article titles and abstracts were used. Copyright statements in Scopus abstracts were removed. Author names and publication years were used for article candidate preselection, as described next.

At this stage, the validation data set consists of all true positive and a limited number of true negative pairs of thesis records and article records. Yet our envisioned PhD candidate article identification system in principle must be able to identify the right article records among all records in the Scopus database. As it would be too computationally costly to actually compare each thesis record with all article records, we have created an heuristic candidate article prefilter method for the Scopus data as part of our larger article identification system. Because this procedure is only of minor importance to this study, its description is deferred to Appendix A1 in the supplemental material. This filtering stage reduces the number of candidates per thesis record to about 1,500 Scopus article records on average.

3.2. Training Data

Two of the methods that we experiment with, LSA and RI, require parallel bilingual training data. This means texts in the two languages of the models that are direct translations. As we are working only with texts in the scientific domain and large, manually translated text corpora are not available for this domain for the English-German language pair, we obtained purpose-specific training data. Bilingual scientific texts were collected from the abstracts of journals that provide both German and English abstracts (50,184 abstracts). A second source of bilingual data is dissertation abstracts, which were obtained from universities' publication servers. We collected 30,275 abstracts of doctoral theses from 10 German universities. Furthermore, we used research project descriptions of projects funded by three funding organizations, the German DFG[9], the Swiss SNF (obtained from the P3 database[10]) and the EU ERC (obtained from the CORDIS database[11]). We used 21,609 DFG, 685 SNF, and 4,997 ERC project descriptions.

We also included the German-English dictionary from the BEOLINGUS translation service of Technical University Chemnitz, version 1.8[12], as doing so generally improves retrieval performance compared to parallel text alone (Xu & Weischedel, 2005).

After preprocessing, the bilingual document-level corpus, without the dictionary, had a size of 14.4 million German and 15.8 million English tokens in 108,000 parallel documents.

[9] https://gepris.dfg.de/
[10] https://p3.snf.ch/
[11] https://cordis.europa.eu/
[12] https://dict.tu-chemnitz.de/, https://ftp.tu-chemnitz.de/pub/Local/urz/ding/de-en/

German documents had on average 134 terms, English documents 146 terms. The dictionary contains 190,000 translations. The entire corpus, including the dictionary, contained 800,000 different German and 240,000 different English terms and is therefore large enough for training in cross-lingual information retrieval (Vulić & Moens, 2015; Xu & Weischedel, 2005).

4. METHODS

4.1. Preprocessing

The text data of the bilingual training data and the test data (dissertation titles + keywords and article titles + abstracts) is processed by removing stopwords (R package `stopwords`; Benoit, Muhr, & Watanabe, 2020), tokenizing (including lowercasing) and language-specific stemming (R package `tokenizers`; Mullen, Benoit et al., 2018), removing numeric tokens (R package `tm`; Feinerer, Hornik, & Meyer, 2008), and discarding tokens of one and two characters length[13]. The stemming uses an interface to the `libstemmer` library implementation of Martin Porter's stemming algorithm, which continues to exhibit high performance (Brychcín & Konopík, 2015). Stemming helps in overcoming the vocabulary mismatch problem by reducing related terms to one common stem (see, for example, Tomlinson (2009) for German and English monolingual retrieval) while at the same time it also reduces the size of the vocabulary. This is not universally beneficial for all terms as some unrelated terms can be conflated to the same stem. Stemming thus can improve recall while incurring some loss in precision. For the fastText experiment, the terms are not stemmed, as the fastText word embeddings were built with unstemmed text, but otherwise processed as described.

4.2. Text Similarity Models

A large number of text similarity calculation methods and models have been proposed over the years in the literature. We have chosen three baseline and two state-of-the-art methods based on the following considerations. We employ both simple baselines and advanced methods as we are interested in whether basic methods show sufficiently good performance on this novel task or whether more recently proposed methods rooted in the semantic embedding paradigm can better handle the task. With regard to the specific choice of models, the vector space model (VSM) has been the central paradigmatic approach in the field of information retrieval for decades and is routinely used as a baseline to compare novel methods against. The LSA model is an early but well-studied representative of the semantic embedding family of methods, which was proposed for multilingual retrieval applications. The *n*-gram similarity method was chosen as it is a conceptually distinct approach from the vector space similarity of all the other considered models and it has shown some promise in multilingual retrieval. For state-of-the-art language-aware semantic embedding methods, we have chosen fastText because of its reportedly good results, wide application, and readily available precomputed vector data, while RI was chosen because it can be trained on custom data with little computational cost and thus serves well for studying the impact of using domain-specific training data to construct task-specific models. As our LSA model is also trained on the same data, we also have an opportunity to compare the performance of these two methods given the same training data.

[13] We also experimented with more sophisticated natural language processing by part-of-speech tagging, lemmatization, and extraction of noun phrases. This proved to be too computationally expensive for application to the entire corpus. The question of whether such higher-quality preprocessing can significantly improve results remains an open issue for further research.

4.2.1. Baseline models

4.2.1.1. Vector space model We use the basic VSM (Salton, Wong, & Yang, 1975) as a language-agnostic baseline. In the VSM, documents are represented by weighted term vectors. We apply standard term frequency-inverse document frequency weighting (tf-idf) to mitigate the distorting effects of unequal term occurrence frequencies. Vector representations of any two documents can be compared by several different vector distance or similarity operations and it is not *a priori* clear which one is best for a specific purpose (Aggarwal, Hinneburg, & Keim, 2001). The conventional choice of similarity measure in the VSM is the cosine similarity and Aggarwal et al. (2001) have shown that for L_k norm distance functions with different values of k the choice of $k = 1$ works well. We therefore experiment with cosine similarity and L_1 distance.

4.2.1.2. LSA model As a second baseline model we construct a joint pseudobilingual vector space using LSA from the same preprocessed English-German parallel corpus introduced in Section 3.2. LSA consists of the application of statistical dimension reduction of the document-term matrix to lower dimensionality to obtain a latent semantics vector space in which terms commonly occurring together in a context (here documents) have similar vectors and in which documents sharing many terms have similar vectors in the same vector space (Deerwester et al., 1990). This method is one way to address the vocabulary mismatch problem for short texts. For two texts to be highly similar according to LSA, they need not share any terms; they only need to contain terms that frequently appeared in the same documents in the training data, or more indirectly, they need to contain terms that appeared in documents that contained other terms that frequently co-occurred in the training data. LSA can be applied to multilingual problems by creating combined multilingual pseudodocuments from translated texts (Dumais et al., 1997). This method is not intrinsically multilingual as there is no information contained in the resulting model about which language a term is from. Therefore, terms of identical spelling with different meanings in different languages will inevitably be conflated to one vector representation.

For this experiment, the preprocessing consisted of tokenization, stopword removal, lowercasing and language-specific stemming. New pseudodocuments were created by concatenating the German and English texts of each document in the training data. From these processed document representations, a tf-idf matrix was created with the `text2vec` `R` package (Selivanov, Bickel, & Wang, 2020). This resulted in an $m \times n$ (297,852 × 923,864) sparse document-term matrix M. Truncated Singular Value Decomposition with the `R` package `RSpectra` (Qiu & Mei, 2019) was applied to the matrix to obtain the latent space model with $t = 1{,}000$ dimensions in which $M \approx U\Sigma V^*$, with U an $m \times t$ document by latent factors matrix of left singular values, Σ a $t \times t$ matrix with the t largest singular values on the diagonal used for weighting the latent factors and all other (off-diagonal) elements being 0, and V^* an $n \times t$ term by latent factors matrix of right singular values. In this latent space it is possible to locate all input documents and all input terms. By calculating the position of new documents based on the latent space training terms which are also contained in the new documents as new 1,000-dimensional vectors it is possible to obtain the similarity of any two new documents regardless of language by calculating the cosine between their vector representations. The dimensionality of the latent vector space is a parameter that can be chosen by simply using the first d dimensions of the vector space to find the best performing value. We try parameter values between 100 and 1,000 by increments of 100.

4.2.1.3. Character n-grams Character n-grams are substrings of n consecutive characters length of larger strings. Segmenting texts into sets of n-grams allows calculating subword-level similarities between texts and is therefore another method that can partially overcome

vocabulary mismatch. *N*-gram similarity has shown good results in several cross-language applications, in particular in related languages (e.g., McNamee & Mayfield, 2004; Potthast, Barrón-Cedeño et al., 2011). Cross-language retrieval with *n*-grams might be assumed to work better for scientific text than general text because many technical terms are highly similar or identical across languages, such as names for health disorders, chemical substances, or organisms. We use the trigram ($n = 3$) implementation of the PostgreSQL version 12 module `pg_trgm`[14]. The module's `pg_trgm.similarity()` function returns a value between 0 (no common trigrams) and 1 (identical strings up to trigram permutation) based on the number of shared trigrams. The score is the ratio of the intersection and the union of the unique trigram sets of the two strings (Jaccard index). Similarities with this function are calculated on the preprocessed dissertation and article text data, where the English and German parts have been concatenated, to keep the input texts into all methods constant. Note that the preprocessing has already eliminated some of the language-specific elements, such as inflections and function words. Applying *n*-gram similarity on this already stemmed data can show if there is additional benefit to move further away from the original words used in the texts than the stemming alone does by splitting the stemmed tokens into *n*-gram subtokens.

4.2.2. Language-aware semantic word embedding models

Word embedding models are vector representations of words (or terms) of fixed dimensionality learned from natural language corpora. Unlike the classic VSM, the dimensionality of the vector space in word embedding models is far smaller than the number of different tokens in a corpus and semantically similar words have similar vectors. LSA is one early example of word embedding models. More than one language's words can be represented in a single word space. Such multilingual models can be constructed either by learning simultaneously from parallel translated texts or by aligning pre-existing monolingual models using some external translation data or other alignment data. As there are quite a number of such models (Ruder et al., 2019), we chose two methods that were straightforward to use or implement, did not require external resources or code dependencies, and were known to scale well. Note that, in contrast to LSA, the following two methods do incorporate information about the language of terms and are thus properly multilingual, having different vector representations for terms of identical spelling in different languages.

4.2.2.1. *FastText aligned multilingual models* FastText is a state-of-the-art word embedding method which achieves good results by learning from *n*-gram subword strings, rather than surface word forms, and representing words in the result vectors as the sum of their constituent *n*-grams (Bojanowski, Grave et al., 2017). This enables the method to overcome difficulties arising from word morphology and rare words. The fastText method is derived from the word2vec Skip-gram with Negative Sampling method (Mikolov, Sutskever et al., 2013). We used the precomputed multilingually aligned models released by the authors of (Joulin, Bojanowski et al., 2018)[15], which are trained on Wikipedia in different languages and aligned after training across languages to map all terms into a common vector space. Note that while Wikipedia is a domain-general knowledge source, it does include vast amounts of scientific knowledge. As the current version of the official fastText programming library is no longer compatible with these vectors, we computed document representations in the database as the average of the fastText word vectors looking up only the exactly matching terms. That means that we cannot benefit from the ability of the fastText library to return results for out-of-vocabulary words.

[14] https://www.postgresql.org/docs/12/pgtrgm.html
[15] https://fasttext.cc/docs/en/aligned-vectors.html

Documents are compared by summing the vectors of their respective terms with tf-idf weights, normalizing the result vectors, and calculating the cosine similarity of these aggregate document representations, following the basic method of Vulić and Moens (2015).

4.2.2.2. Random Indexing RI is an incremental word embedding construction method and a direct alternative to LSA (Sahlgren, 2005). Both use dimensionality reduction techniques to reduce the sparse term-document matrix of a training corpus into a smaller and dense real-valued vector space. In contrast to LSA, in RI the whole term-context matrix, which is usually very large and extremely sparse, is never materialized, the dimension reduction is less computationally demanding, and the model is incremental—it can be updated without a complete recomputation when new data is to be added. RI works by first assigning each document a different static *index vector*, which are vectors of specified dimensionality of values in {−1, 0, 1} drawn from a specific random distribution (Li, Hastie, & Church, 2006). In multilingual RI, there is also a single index vector for each multilingual document. Next, *context vectors* for each term are created by scanning through all documents. For each term, the index vectors of contexts (documents) in which the term occurs are summed in a single pass through the corpus. In this step, term context vectors for each language are generated separately. This projects both languages' words into the same random-vector space (Sahlgren & Karlgren, 2005). Reflective Random Indexing (RRI) is the iterative indexing of contexts (respectively terms) with previously obtained index vectors of terms (respectively contexts) instead of random vectors (i.e., higher order indexing; Cohen, Schvaneveldt, & Widdows, 2010). This way the model can also learn indirect associations between terms that never co-occur in any documents but which co-occur with terms that co-occur, similar to Second Order Similarity (Cribbin, 2011; Thijs, Schiebel, & Glänzel, 2013). Training was done with simple binary occurrence counting of terms in documents—a term was counted as either present or absent, regardless of frequency within a document. To obtain the similarity of two arbitrary documents, the tf-idf weighted context vectors of their constituent terms are added and normalized and then compared with cosine similarity, following Moen and Marsi (2013) and Vulić and Moens (2015), just as in the other methods in the vector space paradigm. The dimensionality of the vector space is also a parameter in RI. Unlike in LSA, here the entire indexing process must be worked through for each different dimensionality parameter value. We also test values between 100 and 1,000 in increments of 100.

The RI methods were also implemented in PostgreSQL 12. For convenience, all vectors are L_2-normalized. We use only second-order context vectors from RRI.

4.3. Remarks

An important distinction of the tested methods needs to be pointed out. The VSM and trigram methods are unsupervised methods—they do not require any training data and work only with the input texts that are to be compared. LSA, fastText, and RI on the other hand are supervised methods. They require training on a text corpus. The fastText vectors we use are the result of training on Wikipedia articles in multiple languages. LSA and RI were trained specifically for this study on the bilingual training data described in Section 3.2, that is, bilingual scientific texts[16]. In particular, we have chosen to train these methods on whole abstracts (or brief project descriptions) for the domain-specific vocabulary and a dictionary for the domain-general vocabulary rather than only on smaller contexts such as sentences or fixed-size word windows. The reason is that we would like to obtain embeddings primarily optimized for document-level topical similarity rather than word-level similarity. In constrast, fastText uses between one and five surrounding words (Bojanowski et al., 2017).

[16] We make our trained LSA, RI, and RRI models of dimensionality 1000 available in Donner (2021a).

Table 2. Examples of term similarities for three terms in LSA, RI, and RRI models

	LSA		RI		RRI	
Rank	Result term	Cosine sim.	Result term (language)	Cosine sim.	Result term (language)	Cosine sim.
Query term: "gdr"						
1	ddr	0.903	ddr (de)	0.760	ddr (de)	0.992
2	sozialkontroll	0.644	ostdeutsch (de)	0.254	ostdeutsch (de)	0.951
3	grenzpolizei	0.642	sed (de)	0.184	reunif (en)	0.951
4	grenzwach	0.642	institut (en)	0.179	republ (en)	0.950
5	ukba	0.642	sbz (de)	0.178	bundesrepubl (de)	0.949
Query term: "groundwat"						
1	grundwassereintrag	0.939	grundwass (de)	0.664	grundwass (de)	0.996
2	grundwasserverhaltnis	0.931	aquif (en)	0.422	aquif (en)	0.987
3	fahigkeitspass	0.849	grundwasserleit (de)	0.377	grundwasserleit (de)	0.982
4	falkenberg	0.835	grundwasserneubild (de)	0.331	aquif (de)	0.979
5	orthophosphatkonzentration	0.835	recharg (en)	0.316	recharg (en)	0.973
Query term: "stahl"						
1	niedriglegiert	0.987	steel (en)	0.594	steel (en)	0.996
2	marag	0.986	austenit (en)	0.227	stainless (en)	0.978
3	maraging	0.986	austenit (de)	0.222	hochf (de)	0.976
4	martensitaushart	0.985	stainless (en)	0.212	werkstoff (de)	0.975
5	nitrierstahl	0.985	fatigu (en)	0.176	tensil (en)	0.971

Note: The term "gdr" is from the abbreviation for German Democratic Republic, "groundwat" is the stemmed form of groundwater, and "stahl" is from German "Stahl" (steel).

To conclude the presentation of the models, Table 2 illustrates term similarities for three example terms for three supervised models. These impressions confirm that the models can learn enough from the training data to provide related result terms.

For each tested method, 1,728,816 similarity calculations between thesis record representations and prefiltered candidate articles are computed. In not every case can a similarity value be obtained for the supervised methods, namely when either the thesis or the article texts do not contain any of the terms of the training corpus. This happens rarely and the exact figures will be given in the next section. There are between one and 47,654 similarity calculation per thesis, with an average of about 1,500. Very few of these are true positives and many theses have no true positives.

5. RESULTS

5.1. Precision and Recall

The evaluation of results for the thesis-article matches data set is not straightforward. The reason is that, for many theses, there are no matching articles, so no matches ought to be found. Such a situation is difficult to evaluate with classic precision and recall methodology as it

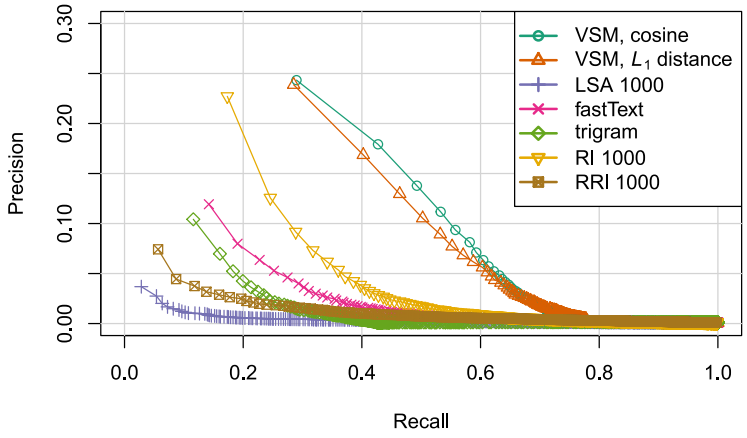

Figure 1. Recall-precision plot.

presuppose true positives for every query. However, we still calculated precision and recall figures to understand the outcomes of this approach despite our reservations. For each evaluated method, up to 1,000 quantile values of the distribution of similarity scores between dissertation text and candidate article text were calculated—fewer if different values actually occurred. The similarity scores at these quantiles were used as threshold values. At each different threshold score, precision and recall were calculated by assigning all document with score greater than or equal to the treshold as positive, those below, negative. We can thus obtain a picture of the possible range of the tradeoff between precision and recall, see Figure 1. Note that here we only report results for LSA, RI, and RRI with dimensionality 1,000 as we found these to be consistently the best values across the tested parameters. Detailed results for the differently parametrized methods can be found in Appendix A2 in the supplemental material.

5.2. Correlation

Another evaluation approach is suggested by the observation that the similarity scores for non-matches (true negatives) should be as low as possible and those for matches (true positives) as high as possible. We construct a new variable by assigning scores of 0 and 1 for true negatives and positives, respectively, and measure the association between this variable and the empirically measured similarity scores of the tested methods with the point-biserial correlation coefficient r_{pb} (Tate, 1954), which is equivalent to the Pearson correlation coefficient when numerical values are assigned to the dichotomous variable. Table 3 shows the averages of the r_{pb} per method weighted by the number of candidates. Note that the absolute values are all very small and they were multiplied by 1,000 for display in the table.

Table 3. Point-biserial correlation between ground truth and similarity methods, multiplied by 1,000

Language	VSM, cosine	VSM, L_1 dist.	trigram	LSA 1000	fastText	RI 1000	RRI 1000	Always 0
de	0.599	0.457	0.534	0.522	0.526	0.548	0.528	0.750
en	0.755	0.299	0.559	0.591	0.549	0.586	0.533	1.039
Combined	0.653	0.402	0.543	0.546	0.534	0.561	0.530	0.850

Table 4. Comparison of standardized similarity scores (*z*-scores)

	VSM, cosine	VSM, L_1 dist.	LSA 1000	fastText	trigram	RI 1000	RRI 1000
Highest scores, theses without publ.	10.11	0.18	2.87	0.99	3.33	2.13	0.85
Theses and associated publ.	11.07	−9.44	1.25	0.59	2.16	2.47	0.95

However, we have reason to doubt that r_{pb} adequately measures the performance we are really interested in. The data set is strongly dominated by true negatives. Due to vocabulary mismatch arising from the very short texts and the bilingual data, the ordinary VSM methods' similarity values are overwhelmingly often exactly 0 (cosine) or 1 (L_1 distance). The other more sophisticated models can compute similarities other than 0 or 1 even if there are no common terms and produce values that are not massively concentrated on one end of the possible range of values. This leads to uninformatively high r_{pb} for the VSM methods. We test this by computing the r_{pb} between a constant (here 0) and the dichotomous match variable. The results are in column "always 0" in Table 3. This method, equivalent to deterministic rejection of any candidate as irrelevant, achieves the best score according to r_{pb}, confirming that the point-biserial correlation coefficient is not a useful evaluation criterion in this particular setting.

5.3. Global Similarity Scores

We have therefore devised the following evaluation method to test how well the scores for the similarity methods can differentiate between constitutive articles of theses and other articles. First, to establish comparability of similarity across methods, the scores for each method are *z*-transformed to obtain scores with mean 0 and standard deviation 1. Score values are expressed as differences from the overall mean in terms of standard deviations. Second, for theses in the sample for which true positives (associated articles) exist, we compute the average of the similarity scores of true positive cases by thesis. Third, for theses without associated articles, we simply select the highest standardized similarity score value. Fourth, we calculate the averages of the scores for the two groups: theses with articles and theses without articles. A good similarity method should have the average similarity value for theses with articles appreciably greater than the average for theses without articles. The results are presented in Table 4, which show that the methods LSA, fastText, and trigram cannot achieve standardized scores for true positives greater than those for most similar articles of theses without associated articles. The two VSM variants and the RI methods exhibit much better performance. In particular, the L_1 distance VSM variant shows more than 9 SD differentiation on average, while the difference of the VSM cosine method is about 1 SD, and those of RI and RRI are 0.3 and 0.1 SD, respectively. Note that the value for L_1 distance VSM for the average standardized similarity is negative because the distance values for similar items are smaller than average values, unlike for the other methods, where the similarities are greater for more similar items.

5.4. Local Similarity Ranks

Another issue is that the density of neighbors in similarity vector space is probably not uniform. If there are more and less dense regions, then the global similarity scores are less informative than the local scores, that is, the scores of similarities of candidates for a single thesis. Consequently, it seems more prudent to look at similarity ranks, stratified by thesis, rather than global similarity values. However, as there are no true positives for theses without constitutive articles, we cannot cover these cases by using this approach.

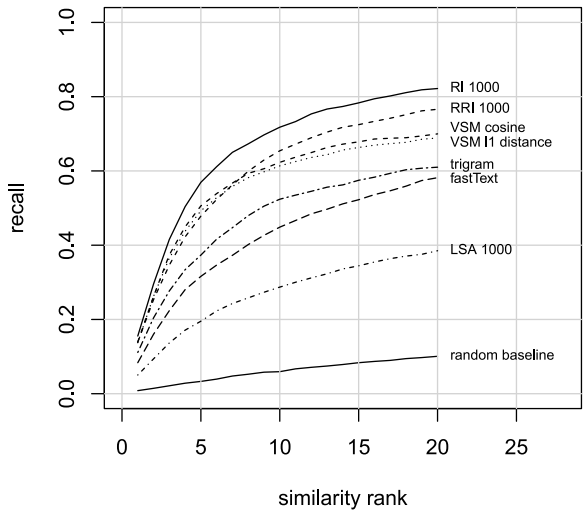

Figure 2. Recall across similarity rank positions.

We proceed with the analysis of recall scores at different rank positions across the methods. Figure 2 shows the curves of the recall values for each considered similarity model for ranks 1 through 20; higher ranks are not interesting as any thesis only has a few integrated articles, if any. Again, these values only include observations of theses that do contain published material, not those that do not. RI shows the best performance here, followed by the baseline VSM methods. RI can achieve 0.8 recall at rank 20 on average, out of some 1,500 candidates per thesis.

To assess if the methods are biased in cross-language similarity measurement we split the validation data and compute recall by rank separately for item pairs with the same language and for pairs with different languages. Figure 3 displays the results. We find that the VSM methods, LSA, and fastText perform much worse in recall if items are of different languages.

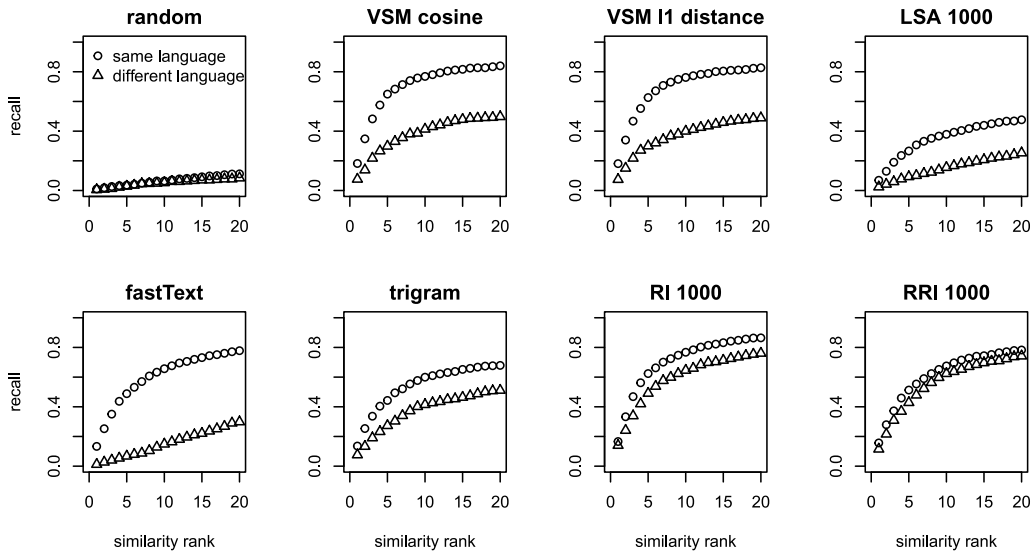

Figure 3. Recall across similarity rank positions by language concordance.

Table 5. Missing similarity values

Total	VSM, cosine	VSM, L_1 dist.	LSA	Trigram	fastText	RI	RRI
1,728,816	3	3	617	3	117	624	624

Trigram is somewhat less affected, RI is modestly affected, and RRI only slightly affected, exhibiting the least cross-language bias.

Finally, Table 5 shows the number of missing similarity values per method. There were three cases in which no method could calculate a score, as the processing of the Scopus texts left no terms to represent the documents. The supervised methods had additional missing values in cases when no terms in the processed text, either thesis or article, were present in the training data. However, the number of missing values is very small compared to the overall number of similarity calculations for all methods.

In summary, these results indicate that, somewhat unexpectedly, the baseline VSM similarity methods perform quite adequately, particularly when considering global similarity values. However, the best performing method when evaluating recall at low ranks is RI, whereas RRI performs a little worse. The pseudo-multilingual baseline method LSA shows only moderate performance, but clearly works to some extent, as can be seen from its results far exceeding random scores. FastText and trigram exhibit intermediate performance.

6. DISCUSSION

Before we proceed to the discussion of the results, a few limitations of this study need to be acknowledged. Because of the large number of choices that can be made in any information retrieval study, it is practically impossible to comprehensively cover all reasonable combinations of methods, settings, parameters, preprocessing steps, and so on. There are many different suggested methods for similarity calculation on vectors, term weighting, vector training, vocabulary pruning, and stop word removal. Parameters for supervised methods such as word context window size or parallel text alignment level could be varied. We have not tested more sophisticated methods for the composition of document-level representations from terms or of preprocessing steps such as decompounding, lemmatization, or word sense disambiguation. All of these factors could influence the results. To keep the scope of the study within feasible limits, we have chosen to apply only basic preprocessing and standard weighting and similarity methods. In the choice of evaluated methods, we have decided in favor of one representative state-of-the-art multilingual word vector method (fastText) and the conceptually attractive but little investigated trigram and RI methods.

ECRs, in particular doctoral students, publish many research outputs. Reliable quantitative estimates of their contribution to the total output of a country have hitherto been elusive, as has the assessment of the scientific impact of their research. Because all graduated PhDs have published a doctoral thesis, we have taken PhD thesis data as the starting point of our approach to quantify doctoral students' research contributions. Cumulative doctoral theses consist of already published material; therefore it is crucial to identify their associated articles to quantify the citation impact of the doctoral research project as a whole. Moreover, the share of identified associated articles among all of a country's articles can serve as a lower bound of the scientific contribution of doctoral students in terms of published output. Our prospective system for the identification of PhD thesis articles consists of a candidate article prefiltering stage and a subsequent automatic classification of candidate article records into those that are constitutive articles and those that are not. This second stage is anticipated to be accomplished by

supervised machine learning algorithms trained and evaluated on sample data. For this matching of candidate associated articles to doctoral thesis records, not only the author names, authors' institutional affiliations, and publication dates of candidate matches are important criteria but also the topical similarity of the research outputs. A good measurement of topic similarity can prove crucial in overcoming uncertainties in matching due to name ambiguities.

The text similarity calculation in this setting is demanding because of the brevity of the texts, the use of multiple languages, and the specialized scientific vocabulary. This rules out the unvalidated use of off-the-shelf solutions. No prior work in this setting has come to our attention, so this is a novel task. Following up on the call by Glavaš, Litschko et al. (2019), the present study is also an instance of a "downstream evaluation" of cross-language distributional semantics models. To this end we have tested three baseline and two state-of-the-art short text similarity methods on a custom validation data set. We collected the necessary training and evaluation data sets and tested the five methods' performance using evaluation measures adapted for the particularities of the data. While this study used German and English language text data, the findings can be informative for any other combination of two or perhaps more languages. Texts were preprocessed for all methods (except fastText) with language-specific stemming and in all similarity calculations (except trigram), tf-idf weights for terms were used.

Our results show that the long-established vector space model of text similarity measurement exhibits quite good performance for this task, likely benefiting from the fact that for one of the texts to be compared (Scopus article records) there will always be some English text and from the specialized scientific terminology. Once we look at the ranking results on the level of matching to individual theses, the limitations of the VSM become apparent as the RI method performs clearly better. The multilingual application of RI has so far only received limited attention (Fernández, Esuli, & Sebastiani, 2016; Moen & Marsi, 2013; Sahlgren & Karlgren, 2005) but the present results are very encouraging. The findings also indicate that the trigram and fastText methods perform moderately well, while LSA is not competitive for this particular task. All methods suffer from some bias when the languages of the compared items differ, but to very different degrees, with RRI being almost unaffected. In conclusion, for the anticipated task of using text similarity as one of a set of features for identifying cumulative theses' associated articles, a combination of VSM cosine similarity score and RI rank can be recommended, with the proviso that the VSM method is by its nature biased in favor of same-language texts. In addition, we can recommend the use of document records of cumulative doctoral theses and their constitutive articles as benchmark data sets for cross-language short text similarity tasks.

ACKNOWLEDGMENTS

The author would like to thank the Information Management group of Deutsche Forschungsgemeinschaft for providing the bilingual project description data of funded DFG projects and Beatrice Schulz for her help in data collection. This research has made use of the PostgreSQL database system and the contributed extensions `aggs_for_vecs` and `floatvec`.

FUNDING INFORMATION

Funding was provided by the German Federal Ministry of Education and Research (grant numbers 01PQ16004 and 01PQ17001).

COMPETING INTERESTS

The author has no competing interests.

DATA AVAILABILITY

Data is made available at https://doi.org/10.5281/zenodo.4733850 and https://doi.org/10.5281/zenodo.4467633 except for proprietary data from Elsevier Scopus. Programming code is made available at https://gitlab.com/pdonner/ri_sql.

REFERENCES

Adrian, D., Ambrasat, J., Briedis, K., Friedrich, C., Fuchs, A., … Wegner, A. (2020). *National Academics Panel Study (Nacaps) 2018* [Data set]. https://doi.org/10.21249/DZHW:nac2018:1.0.0

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In J. Van den Bussche & V. Vianu (Eds.), *International conference on database theory 2001* (pp. 420–434). https://doi.org/10.1007/3-540-44503-X_27

Benoit, K., Muhr, D., & Watanabe, K. (2020). *stopwords: Multilingual stopword lists*. Retrieved from https://CRAN.R-project.org/package=stopwords

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.1162/tacl_a_00051

Brandt, G., Briedis, K., de Vogel, S., Jaksztat, S., Kovalova, I., … Teichmann, C. (2020). *DZHW PhD Panel 2014* [Data set]. https://doi.org/10.21249/DZHW:phd2014:4.0.0

Brauer, J., Oelsner, K., & Boelter, S. (2019). Der wissenschaftliche Nachwuchs in Deutschland: Die Erfassung von Promovierenden und Promovierendendaten. *Dokumentation der Jahrestagung 2019 der GfHf*. Retrieved from https://www.gfhf2019.de/api-v1/article/!/action/getPdfOfArticle/articleID/3048/productID/34/filename/article-id-3048.pdf

Brychcín, T., & Konopík, M. (2015). HPS: High precision stemmer. *Information Processing & Management*, *51*(1), 68–91. https://doi.org/10.1016/j.ipm.2014.08.006

Cohen, T., Schvaneveldt, R., & Widdows, D. (2010). Reflective random indexing and indirect inference: A scalable method for discovery of implicit connections. *Journal of Biomedical Informatics*, *43*(2), 240–256. https://doi.org/10.1016/j.jbi.2009.09.003, PubMed: 19761870

Consortium for the National Report on Junior Scholars. (2017). *2017 national report on junior scholars. Statistical data and research findings on doctoral students and doctorate holders in Germany. Overview of Key Results*. Retrieved from https://www.buwin.de/dateien/buwin-2017-keyresults.pdf

Cribbin, T. (2011). Discovering latent topical structure by second-order similarity analysis. *Journal of the American Society for Information Science and Technology*, *62*(6), 1188–1207. https://doi.org/10.1002/asi.21519

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Donner, P. (2021a). *Bilingual English-German word embedding models for scientific text* [Data set]. https://doi.org/10.5281/zenodo.4467633

Donner, P. (2021b). *Ground truth data for "Identifying publications of cumulative dissertation theses by bilingual text similarity"* [Data set]. https://doi.org/10.5281/zenodo.4733850

Dumais, S. T., Letsche, T. A., Littman, M. L., & Landauer, T. K. (1997). Automatic cross-language retrieval using latent semantic indexing. *AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*, *15*, 21.

Echeverria, M., Stuart, D., & Blanke, T. (2015). Medical theses and derivative articles: Dissemination of contents and publication patterns. *Scientometrics*, *102*(1), 559–586. https://doi.org/10.1007/s11192-014-1442-0

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, *25*(5). https://doi.org/10.18637/jss.v025.i05

Fernández, A. M., Esuli, A., & Sebastiani, F. (2016). Lightweight random indexing for polylingual text classification. *Journal of Artificial Intelligence Research*, *57*, 151–185. https://doi.org/10.1613/jair.5194

Fräßdorf, A., & Fräßdorf, M. (2016). *Is there a doctor on board? Collecting generalizable data on doctoral candidates in Germany* (Discussion Paper No. 1587). Deutsches Institut für Wirtschaftsforschung.

Gerhardt, A., Briede, U., & Mues, C. (2005). Zur Situation der Doktoranden in Deutschland–Ergebnisse einer bundesweiten Doktorandenbefragung. *Beiträge zur Hochschulforschung*, *27*(1), 74–95.

Glavaš, G., Litschko, R., Ruder, S., & Vulić, I. (2019). How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 710–721). https://doi.org/10.18653/v1/P19-1070

Hähnel, S., & Schmiedel, S. (2016). *Promovierende in Deutschland: Wintersemester 2014/2015*. Statistisches Bundesamt.

Heinisch, D. P., & Buenstorf, G. (2018). The next generation (plus one): An analysis of doctoral students' academic fecundity based on a novel approach to advisor identification. *Scientometrics*, *117*(1), 351–380. https://doi.org/10.1007/s11192-018-2840-5

Heinisch, D. P., Koenig, J., & Otto, A. (2020). A supervised machine learning approach to trace doctorate recipients' employment trajectories. *Quantitative Science Studies*, *1*(1), 94–116. https://doi.org/10.1162/qss_a_00001

Joulin, A., Bojanowski, P., Mikolov, T., Jégou, H., & Grave, E. (2018). Loss in translation: Learning bilingual word mapping with a retrieval criterion. *Proceedings of the 2018 conference on empirical methods in natural language processing*. https://doi.org/10.18653/v1/D18-1330

Kenter, T., & De Rijke, M. (2015). Short text similarity with word embeddings. *Proceedings of the 24th ACM international on conference on information and knowledge management* (pp. 1411–1420). https://doi.org/10.1145/2806416.2806475

Koopman, R., Wang, S., & Englebienne, G. (2019). Fast and discriminative semantic embedding. *Proceedings of the 13th international conference on computational semantics* (pp. 235–246). https://doi.org/10.18653/v1/W19-0420

Landauer, T. K., Laham, D., & Derr, M. (2004). From paragraph to graph: Latent semantic analysis for information visualization. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5214–5219. https://doi.org/10.1073/pnas.0400341101, PubMed: 15037748

Li, P., Hastie, T. J., & Church, K. W. (2006). Very sparse random projections. *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 287–296). https://doi.org/10.1145/1150402.1150436

McNamee, P., & Mayfield, J. (2004). Character *n*-gram tokenization for European language text retrieval. *Information Retrieval*, *7*(1–2), 73–97. https://doi.org/10.1023/B:INRT.0000009441.78971.be

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111–3119.

Mitesser, O. (2008). *Latente semantische analyse zur messung der diversität von forschungsgebieten* (Master's thesis, Humboldt-Universität zu Berlin, Philosophische Fakultät). https://doi.org/10.18452/18236

Mitesser, O., Heinz, M., Havemann, F., & Gläser, J. (2008). Measuring diversity of research by extracting latent themes from bipartite networks of papers and references. *Proceedings of WIS 2008, Berlin. Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting*.

Moen, H., & Marsi, E. (2013). Cross-lingual random indexing for information retrieval. *International conference on statistical language and speech processing* (pp. 164–175). https://doi.org/10.1007/978-3-642-39593-2_15

Mullen, L. A., Benoit, K., Keyes, O., Selivanov, D., & Arnold, J. (2018). Fast, consistent tokenization of natural language text. *Journal of Open Source Software*, *3*(23), 655. https://doi.org/10.21105/joss.00655

Natale, F., Fiore, G., & Hofherr, J. (2012). Mapping the research on aquaculture. A bibliometric analysis of aquaculture literature. *Scientometrics*, *90*(3), 983–999. https://doi.org/10.1007/s11192-011-0562-z

Potthast, M., Barrón-Cedeño, A., Stein, B., & Rosso, P. (2011). Cross-language plagiarism detection. *Language Resources and Evaluation*, *45*(1), 45–62. https://doi.org/10.1007/s10579-009-9114-z

Qiu, Y., & Mei, J. (2019). *RSpectra: Solvers for large-scale eigenvalue and SVD problems*. Retrieved from https://CRAN.R-project.org/package=RSpectra

Ruder, S., Vulić, I., & Søgaard, A. (2019). A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, *65*, 569–631. https://doi.org/10.1613/jair.1.11640

Sahlgren, M. (2005). An introduction to random indexing. *Methods and applications of semantic indexing workshop at the 7th international conference on terminology and knowledge engineering, August 16, 2005, Copenhagen, Denmark*.

Sahlgren, M., & Karlgren, J. (2005). Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Natural Language Engineering*, *11*(3), 327. https://doi.org/10.1017/S1351324905003876

Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620. https://doi.org/10.1145/361219.361220

Selivanov, D., Bickel, M., & Wang, Q. (2020). *text2vec: Modern text mining framework for R*. Retrieved from https://CRAN.R-project.org/package=text2vec

Senatskommission für Klinische Forschung, Deutsche Forschungsgemeinschaft. (2010). *Strukturierung der wissenschaftlichen Ausbildung für Medizinerinnen und Mediziner*. Retrieved from Deutsche Forschungsge-meinschaft website: https://www.dfg.de/download/pdf/dfg_im_profil/geschaeftsstelle/publikationen/medizinausbildung_senat_klinische_forschung.pdf

Tate, R. F. (1954). Correlation between a discrete and a continuous variable. Point-biserial correlation. *The Annals of Mathematical Statistics*, *25*(3), 603–607. https://doi.org/10.1214/aoms/1177728730

Thijs, B., Schiebel, E., & Glänzel, W. (2013). Do second-order similarities provide added-value in a hybrid approach? *Scientometrics*, *96*(3), 667–677. https://doi.org/10.1007/s11192-012-0896-1

Tomlinson, S. (2009). German, French, English and Persian retrieval experiments at CLEF 2009. *Working notes for CLEF 2009 workshop*. Citeseer.

Vulić, I., & Moens, M.-F. (2015). Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval* (pp. 363–372). https://doi.org/10.1145/2766462.2767752

Wang, S., & Koopman, R. (2017). Clustering articles based on semantic similarity. *Scientometrics*, *111*(2), 1017–1031. https://doi.org/10.1007/s11192-017-2298-x

Wissenschaftsrat. (2014). *Empfehlungen zu forschungs-und lehrförderlichen Strukturen in der Universitätsmedizin* (No. 5913-04). Retrieved from Wissenschaftsrat website: https://www.wissenschaftsrat.de/download/archiv/5913-04.pdf

Xu, J., & Weischedel, R. (2005). Empirical studies on the impact of lexical resources on clir performance. *Information Processing & Management*, *41*(3), 475–487. https://doi.org/10.1016/j.ipm.2004.06.009

Zamudio Igami, M. P., Bressiani, J. C., & Mugnaini, R. (2014). A new model to identify the productivity of theses in terms of articles using co-word analysis. *Journal of Scientometric Research*, *3*(1), 3–14. https://doi.org/10.4103/2320-0057.143660