OpenMSD: Towards Multilingual Scientific Documents Similarity Measurement

Yang Gao*, **Ji Ma***, **Ivan Korotkov***, **Keith Hall**†, **Dana Alon*** **and Donald Metzler***
* Google Research, † Sizzle AI
{gaostayyang, maji, ivankr, danama, metzler}@google.com
khallbobo@gmail.com

Abstract

We develop and evaluate multilingual *scientific documents similarity measurement* models in this work. Such models can be used to find related papers in different languages, which can help multilingual researchers find and explore papers more efficiently. We propose the first multilingual scientific documents dataset, *Open-access Multilingual Scientific Documents* (OpenMSD), which has 74M papers in 103 languages and 778M citation pairs. With OpenMSD, we develop multilingual SDSM models by adjusting and extending the state-of-the-art methods designed for English SDSM tasks. We find that: **(i)** Some highly successful methods in English SDSM yield significantly worse performance in multilingual SDSM. **(ii)** Our best model, which enriches the non-English papers with English summaries, outperforms strong baselines by 7% (in mean average precision) on multilingual SDSM tasks, without compromising the performance on English SDSM tasks.

Keywords: Scientific Documents, Multilingual Resources, Similarity Measurement

1. Introduction

Although English is the predominant language in scientific publications (Liu, 2017), *diversity and internationalization* in the scientific community has attracted more attention in recent years (Uzuner, 2008; Márquez and Porras, 2020). Over 75% of researchers use English as a foreign language (Baskaran, 2016), and they often need to search related papers in both their native languages and in English. Think tanks and decision-making agencies also need to find related works in different languages on the same topic, e.g., natural resource management and biodiversity studies, to ensure their analyses and decisions are unbiased and consider all affected countries (Steigerwald et al., 2022). As the volume of non-English papers has rapidly grown since 2000, steadily accounting for 5-10% of all scientific publications (Fortunato et al., 2018; Bornmann et al., 2021; Moskaleva and Akoev, 2019), the scientific community has an ever stronger need for multilingual *scientific documents similarity measurement* (SDSM) models, so as to help researchers find, discover, and explore scientific publications in different languages more efficiently. This paper focuses on the development and evaluation of *multilingual SDSM* models.

The state-of-the-art SDSM models, e.g, (Cohan et al., 2020; Ostendorff et al., 2022; Mysore et al., 2022), use Transformer-based (Vaswani et al., 2017) text encoders to create dense representations for the papers. Starting from a *pretrained science-specialized language model* (e.g., SciBERT (Beltagy et al., 2019)), they fine-tune a *dual encoder* model (Gillick et al., 2018) with *contrastive learning objectives* (Chopra et al., 2005; Wu et al., 2018), by using "related" and "unrelated" pairs of papers derived from citation-based heuristics or *graph embedding* algorithms (Perozzi et al., 2014; Lerer et al., 2019). These models show promising performance on several SDSM tasks, e.g., citation prediction. However, all of these SDSM models are trained with English data and hence only work for English papers.

In this paper, we propose both data and novel methods for the *multilingual SDSM* problem. For data, we build the *Open-access Multilingual Scientific Documents* (OpenMSD) dataset, with 74M papers and 778M citations. Key statistics of OpenMSD are presented in Table 1. Three SDSM tasks – *citation*, *co-citation* (Small, 1973), and *bibliographic-coupling* (Kessler, 1963) prediction – are derived from OpenMSD. Compared to *S2ORC* (Lo et al., 2020), a widely used English-only scientific documents dataset, OpenMSD has a comparable number of papers (74M in OpenMSD vs 81M in S2ORC) but 3x more full-content papers (38M in OpenMSD vs 12M in S2ORC) and 2x more citation pairs (759M in OpenMSD vs 381M in S2ORC). To the best of our knowledge, OpenMSD is the first multilingual scientific documents and citation relations dataset. Due to copyright and license restrictions, we cannot directly release the OpenMSD dataset, but publish the scripts for constructing the dataset at https://github.com/google-research/google-research/tree/master/OpenMSD.

With OpenMSD, we develop and evaluate multilingual SDSM models. We derive a training set from OpenMSD and use it to train two state-of-the-art SDSM methods, *Specter* (Cohan et al., 2020) and *SciNCL* (Ostendorff et al., 2022), and test their performance in both English-only (SciDocs, Cohan

LREC-COLING 2024, pages 12467–12480

Papers (74M)	#Papers	
	• w/ abstracts	74M
	• w/ citations	53M
	• w/ full content	38M
	• in English	65M
	#Abstract avg tokens	288
	#Content avg tokens	5448
	#Total tokens	228B
	#Languages	103
	#Categories	340
Citation Pairs (778M)	#En→En	759M
	#En→nonEn	6M
	#nonEn→En	11M
	#nonEn→nonEn	2.5M

Table 1: Key statistics of the OpenMSD dataset.

et al., 2020) and multilingual (test set of OpenMSD) SDSM tasks. Results suggest that, while Specter yields strong performance in both English-only and multilingual SDSM, SciNCL performs poorly in multilingual SDSM, and we find that it is due to the poor performance of the *graph embedding* algorithms (Lerer et al., 2019) on the multilingual citation graphs. To further improve the multilingual performance of Specter, we extend it with several novel methods, including the use of different citation-based heuristics to create training examples, and the use of generative language models to enrich the non-English papers. Our best model significantly outperforms the original Specter in the test set of OpenMSD by 7% (in terms of mean average precision), without compromising the performance on English-only SDSM tasks.

2. Related Works

Scientific documents dataset. Several scientific documents datasets have been compiled with open-access papers. The *arXiv Dataset* (arXiv.org, 2023) contains the metadata and PDFs of 1.7M papers, and the *PMC Open Access Subset* (Bethesda, 2003) contains the full contents of 8M papers from PubMed. Papers on the *ACL Anthology*[1] have also been used to build datasets, e.g., the *ANN dataset* (Radev et al., 2009) with 14K papers and 55K citations, the *ACL ACR dataset* (Bird et al., 2008) with 11K papers, and the *ACL 60-60 dataset* (Diab and Yifru, 2022; Salesky et al., 2023), which provides machine translation of 10K paper titles and abstracts randomly selected from the ACL Anthology from 2017 to 2021, and all the titles and abstracts from ACL 2022 (1.3K) into 60 languages. The Allen AI Institute has published the *S2ORC* dataset (Lo et al., 2020) with 81M papers, the *SciDocs* dataset (Cohan et al., 2020) with over 120K papers and several categories of scientific tasks (classification, SDSM, recommendation), and the *S2AG* API (Kinney et al., 2023), which allows registered users to get access to the metadata (e.g., title, authors, abstract, but no full content) of 206M papers and their citations (2.5B). However, these datasets either lack citations (the ACL-, PubMed- and arXiv-based datasets), or only have English papers (the other mentioned datasets).

Multilingual Language Models. With the success of Transformer-based (Vaswani et al., 2017) language models for English tasks, a number of multilingual variants have also been proposed. They follow the same recipe (e.g., architecture, learning objectives, etc.) as their original English versions, but are pretrained with multilingual texts. Popular variants include the *encoder-only* models like mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020) and mDeBERTa (He et al., 2021), the *encoder-decoder* models like mT5 (Xue et al., 2021) and mBART (Tang et al., 2020), and the *decoder-only* models like XGLM (Lin et al., 2021) and BLOOM (Scao et al., 2022). These models are benchmarked on multilingual datasets like XTREME (Hu et al., 2020) and SuperGLUE (Wang et al., 2019), which include a wide range of tasks like named entity recognition, natural language inference, and question answering. Some of them have also been fine-tuned to tackle downstream science-related tasks, e.g., multilingual acronym extraction in scientific papers (Veyseh et al., 2022), and multilingual bias evaluation in social science papers (Talat et al., 2022). However, there are no multilingual language models specialized for SDSM tasks.

SDSM models. A classic method to measure the similarity and relatedness between papers is *citation analysis* (Zunde, 1971; Nicolaisen, 2007). Based on the citation links between papers, heuristics have been developed to find related papers, e.g., *co-citation* (two papers both cited by some common papers (Small, 1973)), and *bibliographic-coupling* (two papers both cite some common papers (Kessler, 1963)). However, these methods do not work well for papers with sparse citation links, e.g., papers that are newly published, in less-studied topics, or in non-English languages.
Neural-based SDSM methods use different strategies to derive "related" and "unrelated" pairs from the citation relations, and use them to fine-tune science-specialized language models, e.g., SciBERT (Beltagy et al., 2019). For example, Specter (Cohan et al., 2020) uses direct and indirect citation links to derive related unrelated papers, respectively. Ostendorff et al. (2022) proposed the *Scientific documents Neighborhood Contrastive Learning* (SciNCL) method, which uses *graph embedding* algorithms (Lerer et al., 2019) to measure the "distance" of papers in the citation graph, and derive

[1] https://aclanthology.org/

related and unrelated papers based on distance-based heuristics. More details of Specter and SciNCL are presented in §4.2. Mysore et al. (2022) proposed the *Aspire* method, which considers the papers that are co-cited in the same sentence as positive pairs, because close proximity provides a more precise indication of the relatedness of the papers. Furthermore, as the citing sentences typically describe how the co-cited papers are related, they use the citing sentences as an additional signal to guide the model to learn on which aspects the papers are related. However, Aspire requires tools to parse the citations in papers content, which are unavailable for multilingual scientific documents. Also, all these methods are designed for English SDSM; it remains unclear whether they can be used to train multilingual SDSM models.

3. The OpenMSD Dataset

The scientific documents in OpenMSD are extracted from two open-access data sources: *Unpaywall snapshot*[2] (version 202203, with 140M data entries), and *CrossRef Metadata* (Crossref, 2022a) (2022 April snapshot, with 134M data entries). Each data entry includes the title, Digital Object Identifier (DOI), URLs and some additional meta information for a scientific publication. 130 million papers occur in both data sources, by matching the DOIs. We scrape and clean the contents from the URLs, and remove the papers for which no text is extracted; 74M papers are retained, among which 38M have full content.

Citation relations in OpenMSD are extracted from *OpenCitations* (Peroni and Shotton, 2020) (2022 October snapshot). It has 1.4 billion citation pairs, each pair identified by the DOIs of its citing and cited paper. 96% of the DOIs in OpenCitations can be found in Unpaywall or CrossRef. We only keep citation pairs that have both the citing and cited papers' abstract extracted, as papers without abstracts cannot be used to train SDSM models. 778M citation pairs are kept in the end.

We use *cld3*[3] to detect the languages from papers' titles and abstracts. 103 languages are found, with English (65M) being the predominant language, followed by German (2.6M) and French (1.2M). All other languages have fewer than 1M papers, and among them 42 languages have fewer than 100 papers. Fig. 1 shows the the sizes of the top 20 languages. Papers' category labels are extracted from CrossRef Metadata; 76% papers have category labels, and each paper has 1.4 category labels on average. 340 categories are found in total; the size of the top 20 categories are presented in Fig. 2.

[2] https://unpaywall.org/products/snapshot

[3] https://github.com/google/cld3.

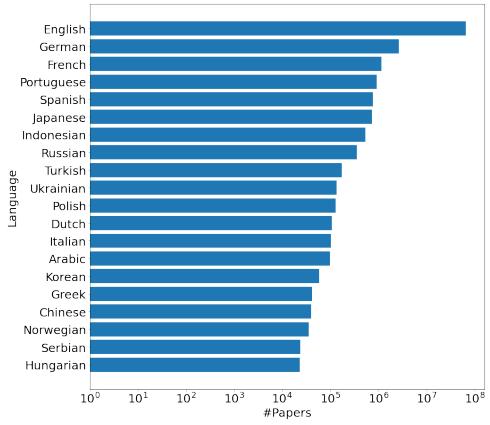

Figure 1: Top 20 languages in OpenMSD.

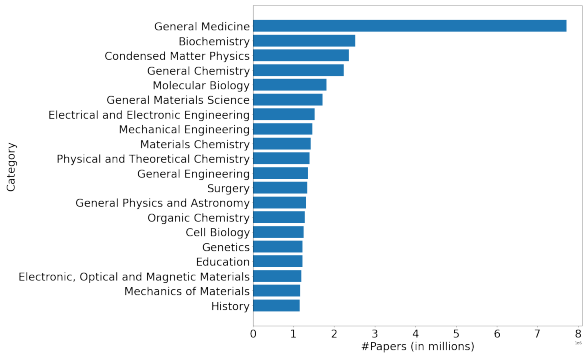

Figure 2: Top 20 categories in OpenMSD.

We note that OpenMSD is dominated by English resources, which account for 88% papers and 98% citation pairs (see Table 1). A common strategy to mitigate the data imbalance is to down-sample the English papers (Conneau et al., 2020), but it only works well in very large datasets like *mC4* (Xue et al. 2021, with 6.6B pages and 6.3T tokens). Some recent works (e.g., (Wang et al., 2022b)) suggest that the English-predominance in the training set does not necessarily hurt the multilingual performance, because fine-tuning multilingual models only with English data can yield strong performance on multilingual tasks. For these reasons, we do not perform any down-sampling over the English resources in OpenMSD. Also, as scientific papers share many common characteristics regardless of their categories, we do not manipulate the category distributions in OpenMSD.

4. Multilingual SDSM

In this section, we decribe how we develop and evaluate multilingual SDSM models. We define the SDSM task in §4.1, describe our SDSM models in §4.2, explain how we split OpenMSD into train and test sets in §4.3, and present the experiment details and results in §4.4.

4.1. Task Definition

The task of SDSM is to find the *related* papers for a query paper. The definition of related is task dependent, and many tasks can be viewed as special cases of SDSM: e.g., in *co-citation prediction*, two papers are related if they are co-cited (i.e., both cited by a third paper); in *co-view prediction* (Cohan et al., 2020), two papers are related if they are often viewed/clicked by the users.

Formally, let $\mathcal{D}$ be a set of documents, and let $R : \mathcal{D} \mapsto 2^{\mathcal{D}}$ be the *relatedness relation* between them: For $p \in \mathcal{D}$, $R(p) \subseteq \mathcal{D}$ is the set of documents that are related to p. The task of SDSM is to learn a similarity measurement function $s : \mathcal{D} \times \mathcal{D} \mapsto \mathbb{R}$, so that for documents $p, q^+, q^- \in \mathcal{D}$, $s(p, q^+) > s(p, q^-)$ if q^+ is a related document for q (i.e., $q^+ \in \mathcal{R}(p)$) and q^- is unrelated (i.e., $q^- \notin \mathcal{R}(p)$). By *multilingual SDSM*, we mean that (i) documents in $\mathcal{D}$ are in different languages, and (ii) the related document $q^+ \in \mathcal{R}(p)$ can be in a different language from the query document p.

4.2. SDSM Models

In this section, we describe the methods to develop the multilingual SDSM models. In particular, we consider two state-of-the-art SDSM methods, *Specter* (Cohan et al., 2020) and *Scientific documents Neighborhood Contrastive Learning* (SciNCL, Ostendorff et al., 2022), and adapt them to the multilingual setup. In addition, we extend Specter by using different strategies to get the training examples.

Specter uses a Transformer-based model (e.g., SciBERT, Beltagy et al., 2019) as the base model and fine-tunes it with the *triplet hinge loss*. Formally, given a triplet (p_i, q_i^+, q_i^-), where p_i is the query paper, $q_i^+ \in R(p_i)$ a positive example to the query, and $q_i^- \notin R(p_i)$ a negative example, the loss function is

$$\mathcal{L}_{TL} = \max\{0, [s(p_i, q_i^-) - s(p_i, q_i^+) + m]\}, \quad (1)$$

where m is a hyper-parameter. Because labeling the positive and negative examples is too expensive, in practice, Specter uses the following heuristics to derive them: if paper A cites paper B, B cites C but A does not cite C, then (A, B) is used as a positive pair while (A, C) is used as a negative pair.

SciNCL uses the same loss function (Eq. (1)) as Specter in fine-tuning. However, instead of using heuristics to derive the positive and negative examples, it uses *graph embedding* models to derive them. They first run a graph embedding algorithm (e.g., BigGraph by Lerer et al. 2019) on the citation graph to learn the embedding for each node (i.e., paper). With the nodes' embeddings, they use fast nearest neighbor search algorithms (e.g., Xiong et al. 2020) to find the top-K neighbors for each node, and extract positive and negative papers with the following strategy: for each paper, its i-th to $(i+n)$-th closest papers are used as positives, while its k-th to $(k+n)$-th closest papers are used as (hard) negatives, where $i, k, n \in \mathbb{N}^+$ are hyper-parameters. With systematic hyper-parameter search, they find that $i = 20$, $k = 2000$ and $n = 5$ yield the best performance. We have explored other hyper-parameters when we re-implement SciNCL, but we find that the ones used in the original work yield the best performance.

Multilingual Adaptation. The original Specter and SciNCL use the (English-only) SciBERT model as the base model. To adapt them to multilingual SDSM, we develop a multilingual science-tailored model as their new base. More specifically, we further pretrain mT5-base (Xue et al., 2021) using *contrastive loss with sampled in-batch negative* (CL) (Henderson et al., 2017). CL encourages the model to push the positive examples closer and the negative examples apart. Formally, let $\{(p_i, q_i)\}_{i=1}^n$ be a training batch with size n, where (p_i, q_i) is the i-th pair of related documents; CL is defined as

$$\mathcal{L}_{CL} = \frac{-\exp[s(p_i, q_i)]}{\sum_{j=1}^n \exp[s(p_i, q_j))]}. \quad (2)$$

To construct the training example pairs (p_i, q_i), we randomly extract snippets from the documents in the train set of mC4 (Xue et al., 2021) and OpenMSD. Snippets from the same document are treated as positive pairs. The length of each snippet is between 10 and 256 mT5-sentence-piece tokens. CL is used because it showed strong performance in both pretraining (Lee et al., 2019a) and fine-tuning (Giorgi et al., 2021; Izacard et al., 2022) dense representations for information retrieval tasks. We call the new base model *mT5CL*. With mT5CL, we develop the multilingual versions of Specter and SciNCL by following their original fine-tuning recipe but run the training on the train set of OpenMSD. We call the the resulting multilingual models *Multilingual Specter (mSpt)* and *Multilingual SciNCL (mNCL)*, respectively.

Generalized Specter. Specter uses direct citations (DCs) to extract positive pairs. But there are other citation relations, e.g., co-citation (CC) and bibliographic-coupling (BC), widely used as indicators for related documents (see §2). We extend the original Specter method by using a mixture of different citation relations to extract positive training examples:

- **Use the union of DC, CC, and BC pairs.** For example, we can use both DC and CC pairs as positives, denoted as $DC \cup CC$. When the numbers of different types of pairs are not the same, we down-sample the over-represented relations so as to have the same number of pairs from each relation type.

Split	Languages
Train (53M)	En, De, Fr, Ja, Es, Pt, Tr, Ru, Id, It, Nl, Pl, Uk, Ko, Nn, Zh, Cs, Hu, Lt, Da, Sv, Hr, Af, Ms, Vi, Sl, Fi, Ro, Ar, Gl,
Test (85K)	En, Sr, De, Fr, He, Es, Pt, Ja, Fa, Ca, Lv, Tr, La, Sk, Su, Zh, Ru, It, Eu, Pl, Nl, Id, Et, Ko, Cs, Bg, Hu, Sq, Is, No, Hi, Uk, Tl, Az, Af, Lt, Bs, Hr, Ms, Sv, Be, Da, Eo, Mi, Oc, Vi, Cy, Fi, Ia, Kk, Ku, Mk, Ro, Sl, Gl, Ga, Aa, Co, Fo, Ka, El, Ky, Sw, Th, Uz

Table 2: Languages (ISO 639-1 code) in different splits of OpenMSD, ordered by their sizes in each split.

- **Use the intersection of DC, CC, and BC pairs**. Suppose a paper A cites paper B and they are both cited by another paper C, then (A, B) is both a DC and CC pair. We assume that the pairs falling into multiple relation types at the same time have higher similarity level, compared to the pairs that only fall into one relation type. We consider all (four) possible intersection combinations of the relation pairs to build positive pairs: $DC \cap CC$, $DC \cap BC$, $CC \cap BC$, and $DC \cap CC \cap BC$.

The negatives are extracted with the same heuristics as in standard Specter. We use subscript to denote which pairs are used to train mSpt: e.g., *mSpt*$_{DC \cup CC}$ denotes the mSpt model trained with the union of the DC and CC pairs as the positive examples.

4.3. Data Preparation

To use OpenMSD to train and evaluate multilingual SDSM models, we first remove all papers that do not have citation links with any other papers, as we cannot find their "related" papers; this leaves us with 53M papers in 65 languages. The remaining papers are split into train (53M papers) and test (85K papers) sets. The languages in each split are presented in Table 2. Note that we deliberately exclude some languages from the train set and only present them in the test set; this allows us to use the test set to benchmark the models' performance for languages unseen during training.

With the data splits, we derive three types of related paper pairs in each data split: *direct citations*, *co-citations* and *bibliographic-coupling*. We remove pairs between papers across different splits to avoid data leakage. Also, we remove all English to English pairs in the test set, to make sure that the test set is focused on pairs involving non-English papers. The numbers of mono-lingual and cross-lingual pairs of each relation type and in each data split are presented in Table 3.

		Train	Test
DC	#En→En	759M	0
	#En→nonEn	6M	3K
	#nonEn→En	11M	6K
	#nonEn→nonEn	3M	3K
CC	#En↔En	12B	0
	#En↔nonEn	208M	1K
	#nonEn↔nonEn	21M	1K
BC	#En↔En	63B	0
	#En↔nonEn	1B	7K
	#nonEn↔nonEn	29M	1K

Table 3: Sizes of direct citation (DC), co-citation (CC) and bibliographic-coupling (BC) pairs in each data split. Note that DC is a directed relation (denoted by $\rightarrow$), while CC and BC are non-directional relations (denoted by $\leftrightarrow$).

4.4. Experiments

Implementation details. When representing a paper, we use the concatenation of the title and the abstract of the paper as input (in line with Specter and SciNCL), and apply average-pooling to the output of the top transformer layer to get the vector representation. Document similarities are measured by the dot product of the document embeddings. To find the optimal hyper-parameters, we have used batch sizes 256, 512, 1K, 2K and 4K, and initial learning rates 10^{-n}, where $n = 1, 2, \cdots, 7$. The inverse square-root learning rate decay strategy is used, with decay factor 5×10^{-5}, and the minimum learning rate is set to 10^{-8}. We find that batch sizes $\geq 1K$ yield similar performance, and learning rate 10^{-2} yields the best performance on the dev set (0.5% data randomly sampled from the train set). Each model is fine-tuned for up to 100K steps, in which the first 1.5K steps are used for warm-up. Checkpoints with the best performance on the dev set are used at test time. All our experiments are performed on a cloud machine with eight TPUv3s.

Performance on English-Only SDSM. We first test the performance of different models on the (English-only) SciDocs test set. In addition to mSpt and mNCL, we consider the following two categories of models as baselines. (i) Pretrained language models, including SciBERT, mT5 and our newly-prerained mT5CL. (ii) The original (English-only) Specter and SciNCL models, with implementations downloaded from their Github repositories. Comparing against the pretrained models allows us to understand whether the fine-tuning strategies in §4.2 are effective or not, and comparing against the original English-only models allows us to test whether the multilingual models can yield comparable performance on the English-only SDSM tasks.

The performance is measured by the mean average precision (MAP) and nDCG@10 scores.

We consider four tasks in SciDocs: *cite*, *co-cite*, *co-view* and *co-read*. Each task has a pool of 30K papers, grouped into 1K clusters. Each cluster has one query paper, five positive examples (e.g., in the co-view task, a positive example is a paper that is often co-viewed with the query paper) and 25 (randomly sampled) negative examples. However, existing evaluations on SciDocs (Cohan et al., 2020; Ostendorff et al., 2022; Mysore et al., 2022) have shown the *ceiling effect*: their nDCG performance are all 90%+ and the gaps between different methods are rather small (less than 2 percentage points). This is because papers are organized into clusters at test time: given a query paper, models only need to find the five positives from the cluster (30 papers). This setup over-simplifies the SDSM problem, because in practice, the models often need to find the related papers from the a large pool of papers (used as a retriever), or a smaller pool with highly relevant papers (used as a re-ranker). Hence, to make the SciDocs evaluation more realistic, we merge the paper pools of all four tasks and ignore the clusters at test time; i.e., for each query paper, the models need to rank all 120K papers in the merged paper pool to find the five positives.

The results on the merged test set of SciDocs are presented in Table 4. We make the following observations. (i) Among the pretrained language models, mT5CL significantly[4] outperforms the other two pretrained language models, suggesting that further pretraining the models using the CL objective with scientific documents can greatly benefit the SDSM performance. (ii) Average performance of all the mSpt models are significantly better than all the pretrained language models, suggesting that the positive/negative examples extracted with the strategies described in §4.2 are effective. (iii) Comparing the English-only and multilingual models, we find that the multilingual models are generally worse than their English-only counterparts. However, while the gap between mNCL and SciCNL is statistically significant, the average performance gap between the best mSpt models (i.e., all mSpt models with DC in the subscript) and the original Specter is not statistically significant. This observation suggests that the multilingual models can yield comparable performance on the English-only SDSM tasks.

Performance on Multilingual SDSM. The performance of different models on the OpenMSD test set is presented in Table 5. Note that the performance of SciBERT, (the original versions of) Specter and SciNCL are not reported in the table, because they only work for English papers and hence cannot be applied to the OpenMSD test set.

From the results, we make the following main observations. (i) All the mSpt models outperform all the pretrained language models, confirming again the effectiveness of the Specter-related strategies presented in §4.2. (ii) The average performance of all mSpt models is significantly better than the pretrained models; however, mNCL only marginally (and not significantly) outperforms the pretrained models. This observation confirms the success of mSpt but the ineffectiveness of mNCL. We will investigate the failure of mNCL in §4.5. (iii) Among the mSpt models, the versions using DC perform significantly better than the versions without DC. In particular, the version using both DC and CC as positives (i.e., $DC \cup CC$) yields the best performance, better than using of any of the relation types alone or together. This observation suggests that DC is the most effective heuristic for deriving positive examples, and using DC with other types of relations can further improve the performance. This finding is significant as existing works only use DC (Cohan et al., 2020) or CC (Mysore et al., 2022) pairs as positive training examples, but never consider their combinations.

4.5. Investigate mNCL

We note that mNCL performs significantly worse than the mSpt models in both English-only (SciDocs) and multilingual (OpenMSD) SDSM tasks. Because mNCL uses graph embeddings-induced rankings to derive the training examples (see §4.2), we look into the quality of the graph embedding rankings to better understand why mNCL fails.

In particular, we run a popular graph embedding algorithm, BigGraph (Lerer et al., 2019), on the train set of OpenMSD to train the graph embedder model, and apply it to the test set of OpenMSD. Table 6 presents the quality of the rankings derived from the graph embedders with different embedding dimensions. We find that (i) With larger embedding dimension, the embedding models performance increase (note that the original SciNCL work only uses graph embeddings up to 768 dimensions). (ii) However, even with dimension size 2048 (the largest dimension size we can run in reasonable time), the graph embedding's average performance is worse than most of the mSpt models (in MAP; see Table 5). The poor performance of the graph embeddings will yield low-quality positive/negatives used in mNCL, hence harming performance of mNCL.

We believe the above results are important, because (i) they are the first results of applying graph embedding algorithms to a multilingual citation graph, (ii) they show that methods work well in English SDSM tasks may fail in multilingual SDSM.

[4]We use double-tailed t-test $p < 0.05$ as the significance test throughout this paper.

Method	Citation		Co-citation		Co-read		Co-view		Average	
	MAP	nDCG	MAP	nDCG	MAP	nDCG	MAP	nDCG	MAP	nDCG
Pretrained Language Models										
SciBERT	0.52	2.48	0.51	2.58	0.91	3.60	1.21	5.26	0.79	3.48
mT5	0.31	1.41	0.47	2.31	0.49	1.99	0.92	3.89	0.55	2.40
mT5CL	**4.21**	**14.53**	**3.85**	**12.71**	**4.69**	**13.12**	**6.23**	**16.96**	**4.75**	**14.33**
English-only Models										
Specter	5.51	15.96	**5.76**	**16.22**	6.96	**17.32**	**9.47**	**22.43**	6.93	**17.98**
SciNCL	**5.92**	**18.04**	5.62	14.37	**7.56**	16.72	9.24	21.65	**7.09**	17.70
Multilingual Models										
mNCL	3.92	12.74	4.17	12.11	5.27	13.78	7.21	18.05	5.14	14.17
mSpt$_{DC}$	5.60	16.88	5.61	14.87	**7.06**	16.67	9.16	22.22	6.86	17.66
mSpt$_{CC}$	4.75	14.66	5.43	14.43	6.42	14.94	8.59	20.12	6.30	16.04
mSpt$_{BC}$	4.18	13.48	4.37	13.17	5.84	14.35	7.13	17.56	5.38	14.64
mSpt$_{DC\cup CC}$	5.57	16.75	**5.80**	14.57	**7.06**	16.27	**9.23**	22.40	**6.92**	17.50
mSpt$_{DC\cup BC}$	**5.86**	17.31	5.68	**15.27**	6.97	16.26	9.06	21.81	6.89	17.66
mSpt$_{CC\cup BC}$	4.41	13.62	5.00	14.06	6.31	14.77	8.33	19.73	6.01	15.55
mSpt$_{DC\cup CC\cup BC}$	5.56	17.03	5.29	14.05	6.98	16.37	8.66	21.38	6.62	17.21
mSpt$_{DC\cap CC}$	5.63	**17.55**	5.55	14.89	7.00	**16.75**	8.92	**22.75**	6.78	**17.99**
mSpt$_{DC\cap BC}$	5.67	17.23	5.30	14.82	6.28	15.69	9.19	22.69	6.61	17.61
mSpt$_{CC\cap BC}$	4.76	14.72	5.05	14.65	6.65	15.03	8.16	20.13	6.16	16.13
mSpt$_{DC\cap CC\cap BC}$	5.52	17.25	5.34	14.78	6.61	16.04	8.70	21.57	6.54	17.41
mSpt$_{DC\cup CC}$ + English Summaries										
TopNSumm$_{64}$	5.40	16.36	**5.79**	**15.84**	7.31	16.69	9.32	22.08	6.96	17.74
PaLM2Summ$_{64}$	5.41	**16.40**	5.75	15.80	7.26	16.74	9.33	22.07	6.94	17.75
TopNSumm$_{128}$	5.49	16.38	5.77	15.73	7.44	16.94	**9.35**	22.38	7.01	17.86
PaLM2Summ$_{128}$	**5.58**	16.37	**5.79**	15.70	**7.55**	**17.13**	9.30	**22.51**	**7.06**	**17.93**

Table 4: Performance (in %) on the test set of the merged SciDocs (English-only) dataset. All results are averaged over 5-10 runs with different random seeds.

More rigorous investigations are required to better understand the reasons, e.g., comparing and analyzing different graph embedding algorithms, investigating the topological structures of the English-only and multilingual citation graphs, etc. It is beyond the scope of this work and we encourage future works on this topic (also see §7).

5. Enrich the Non-English Documents with English Summaries

Because OpenMSD is dominated by English papers and pairs (see §3), models trained with OpenMSD are exposed more to English training examples. We aim to leverage the model's English capabilities to improve its performance on non-English documents. To this end, inspired by the works on *cross-lingual summarization* (Zhu et al., 2019; Wang et al., 2022a), we propose to create English summaries for the non-English papers, and concatenate the summaries to the original (non-English) text to create *enriched documents*.

As there are no cross-lingual scientific documents summarization datasets or models available, we decide to use two *zero-shot* methods to generate English summaries. **(i)** Using the English translation of the top-N tokens as the summary. This is a simple yet strong baseline widely used in summarization (Gao et al., 2020; Bao et al., 2022). **(ii)** Prompting a large generative language model to write English summaries. The generative model we use is *Flan-PaLM2* (Anil et al., 2023) (version *Otter* on Google Cloud API[5]); the instruction-tuned (Flan) version is used because recent works (Zhang et al., 2023) suggests that even smaller Flan-tuned language models can generate high-quality summaries, better than their larger but non-Flan-tuned counterparts. The English summary is then concatenated to the original text in the following format: *Title: {title_text}. Abstract: ({English_summary_text}) {original_abstract_text}*. Note that English papers are not augmented with any summaries.

We consider summaries with two different lengths: 64 and 128 tokens. To get the top-N translation summaries, we simply truncate the translated abstracts to the target lengths. To prompt Flan-PaLM2 to generate summaries, we experiment with a few prompts and finally use two prompts to generate the short and long summaries, respectively: **(i)** *Summarize the passage below with no more than 30 words in English.* **(ii)** *Extract the three most important findings from the passage below, and trans-*

[5] https://cloud.google.com/vertex-ai

Method	Citation		Co-citation		Bib-couple		Average	
	MAP	nDCG	MAP	nDCG	MAP	nDCG	MAP	nDCG
Pretrained Language Models								
mT5	1.62	2.07	0.99	1.73	0.46	0.72	1.02	1.51
mT5CL	7.81	9.45	4.01	5.81	**1.77**	**2.75**	**4.53**	**6.00**
Multilingual Models								
mNCL	7.92	9.62	4.11	6.00	1.76	2.71	4.60	6.11
mSpt$_{DC}$	17.64	21.15	7.15	9.89	3.70	5.33	9.50	12.12
mSpt$_{CC}$	15.21	18.39	6.41	8.92	3.22	4.71	8.28	10.67
mSpt$_{BC}$	11.72	14.35	5.08	7.28	3.00	4.33	6.60	8.65
mSpt$_{DC\cup CC}$	**18.03**	**21.63**	**7.42**	**10.11**	3.65	5.26	**9.70**	**12.33**
mSpt$_{DC\cup BC}$	17.77	21.35	7.09	9.69	**3.73**	5.28	9.53	12.11
mSpt$_{CC\cup BC}$	14.32	17.32	6.13	8.48	3.22	4.70	7.89	10.17
mSpt$_{DC\cup CC\cup BC}$	17.03	20.40	6.75	9.34	3.63	5.20	9.14	11.65
mSpt$_{DC\cap CC}$	17.29	20.84	6.85	9.55	3.51	5.08	9.22	11.82
mSpt$_{DC\cap BC}$	17.51	21.00	6.97	9.69	3.74	**5.42**	9.41	12.04
mSpt$_{CC\cap BC}$	15.19	18.24	6.41	8.75	3.28	4.70	8.29	10.56
mSpt$_{DC\cap CC\cap BC}$	16.87	20.16	6.59	9.37	3.53	5.10	9.00	11.54
mSpt$_{DC\cup CC}$ + English Summaries								
TopNSumm$_{64}$	18.30	22.02	7.23	9.87	3.74	5.35	9.76	12.41
PaLM2Summ$_{64}$	18.85	22.68	7.29	9.97	3.97	5.68	10.04	12.78
TopNSumm$_{128}$	18.55	22.31	7.23	9.88	3.99	5.72	9.92	12.64
PaLM2Summ$_{128}$	**19.46**	**23.40**	**7.64**	**10.53**	**4.05**	**5.81**	**10.38**	**13.24**

Table 5: Performance (in %) on the test set of OpenMSD. All results are averaged over 5-10 runs with different random seeds.

GraphEmbd Dim.	Citation		Co-citation		Bib-couple		Average	
	MAP	nDCG	MAP	nDCG	MAP	nDCG	MAP	nDCG
128	1.29	1.74	0.51	0.95	2.09	3.15	1.30	1.95
256	1.33	1.78	0.75	1.38	2.79	4.06	1.62	2.41
512	2.89	3.96	1.86	3.69	4.26	6.38	3.00	4.68
1024	4.94	6.69	3.52	7.19	5.94	9.65	4.80	7.84
2048	**8.78**	**11.39**	**6.32**	**12.06**	**8.05**	**13.96**	**7.72**	**12.47**

Table 6: Performance (in %) of the BigGraph (Lerer et al., 2019) embeddings on OpenMSD test set, with different dimension sizes. We have also tried DeepWalk (Perozzi et al., 2014) and InstantEmbedding (Postăvaru et al., 2020) and they yield similar performance and trend.

late them to English. The model tends to generate over-length summaries: the average length the summaries generated with the two prompts above are 71 and 138 tokens, respectively. Over-length tokens are removed to get the final summaries.

The enriched documents are used to train and test mSpt$_{DC\cup CC}$, the strongest variant of mSpt. The results of the proposed method on the English-only and multilingual SDSM tasks are presented in the bottom blocks in Table 4 and 5, respectively. Firstly, we find that using the Flan-PaLM2-generated summaries consistently yields better performance than the top-N translation summaries; we believe this is because Flan-PaLM2 considers the whole abstract when generating the summaries, and hence its summaries are more informative and comprehensive than the top-N translation summaries. Secondly, using the enriched documents yields comparable performance as the other fine-tuned models on English SDSM (SciDocs), but yields significantly better performance than all the other considered methods on multilingual SDSM (OpenMSD), improving the performance of mSpt$_{DC\cup CC}$ by 7% in both MAP and nDCG. These results suggest that enriching the non-English papers with high-quality English summaries can significantly improve the multilingual models' performance for papers in non-English and unseen languages.

6. Conclusion

In this work, we proposed both datasets and novel methods for the multilingual *scientific documents similarity measurement* (SDSM) problem. For data, we built *OpenMSD*, the first multilingual scientific documents dataset, and derived three SDSM tasks therefrom. For methods, we adapted some SDSM methods that are highly successful in English SDSM tasks to the multilingual setup, and

found that some of them fail to generalize well to the multilingual SDSM tasks. We also found that enriching non-English papers with English summaries can yield significantly better performance in multilingual SDSM tasks, without compromising the model's performance in English SDSM. We hope this work will facilitate and encourage more future works on multilingual SDSM.

7. Limitations & Future Works

Performance in individual languages. Due to the space limit and the large number of languages in OpenMSD, we did not present the performance of the models in individual languages or in language groups (e.g., high-, medium- and low-resource languages). We leave these more detailed multilingual studies to future work.

Biases in the data. Biases are often observed in large scale text corpus (Blodgett et al., 2020; Hovy and Prabhumoye, 2021). We note that the languages used in the documents in OpenMSD may have biases (e.g., gender or ethnicity bias), and the documents of different languages may have different biases. Studying how the biases in the scientific documents affect the SDSM models' performance is a highly important topic, and we call for more thorough investigations on it.

Graph Embeddings in OpenMSD. In §4.5, we have shown that the average performance of the rankings derived from the graph embeddings is poor, worse than most mSpt models. However, when looking into the performance in three different tasks (DC, CC and BC prediction), we note that the graph embeddings' performance is strong in CC and BC prediction but poor in DC prediction, and this observation is consistent across graph embedding algorithms and graph embedding dimensions. The reasons remain unclear and we call for investigations from the wider research community, including researchers from machine learning and graph theory.

Generative Language Models. In §5, we showed that using generative models to enrich the non-English papers can yield significant performance improvement on multilingual SDSM. This finding opens up many interesting directions yet to be explored, e.g., techniques to create better cross-lingual summaries with the generative models (prompt engineering, few-shot demonstrations, prompt-tuning, and bigger generative language models), and the impact of the summaries quality on the SDSM models' performance.

Diversity of base models. Our proposed models are based on the mT5-Base model (Xue et al., 2021). It would be interesting to investigate how different models sizes (larger or smaller models) and types (e.g., decoder-only models) affect the models performance; we leave it for future work.

8. Bibliographical References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. PaLM 2 Technical Report.

arXiv.org. 2023. arXiv Dataset.

Forrest Sheng Bao, Ruixuan Tu, and Ge Luo. 2022. Docasref: A pilot empirical study on repurposing reference-based summary quality metrics reference-freely. *arXiv preprint arXiv:2212.10013*.

Angathevar Baskaran. 2016. UNESCO science report: Towards 2030. *Institutions and Economies*, pages 125–127.

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

Bethesda. 2003. PMC Open Access Subset. https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist. [Online; accessed 23-June-2023].

Steven Bird, Robert Dale, Bonnie J Dorr, Bryan R Gibson, Mark Thomas Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir R Radev, Yee Fan Tan, et al. 2008. The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *LREC*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III au2, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp.

Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. 2021. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1):1–15.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020b. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Crossref. 2022a. April 2022 public data file from crossref.

Crossref. 2022b. April 2022 Public Data File from Crossref.

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B Hall, and Ming-Wei Chang. 2023. Promptagator: Few-shot dense retrieval from 8 examples. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Mario S Di Bitetti and Julián A Ferreras. 2017. Publish (in English) or perish: The effect on citation rate of using languages other than English in scientific publications. *Ambio*, 46:121–127.

Mona Diab and Martha Yifru. 2022. ACL 2022 D&I Special Initiative: 60-60, Globalization via Localization. https://www.2022.aclweb.org/dispecialinitiative. [Online; accessed 23-June-2023].

Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. 2018. Science of science. *Science*, 359(6379):eaao0185.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *CoRR*, abs/2104.08821.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. Supert: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354.

Daniel Gillick, Alessandro Presta, and Gaurav Singh Tomar. 2018. End-to-end retrieval in continuous space. *arXiv preprint arXiv:1811.08008*.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. Declutr: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 879–895.

Joshua Goodman. 2001a. A bit of progress in language modeling. *CoRR*, cs.CL/0108005v1.

Joshua T. Goodman. 2001b. A bit of progress in language modeling. *Computer Speech & Language*, 15(4):403–434.

R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and linguistics compass*, 15(8):e12432.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Rebecca Hwa. 1999a. Supervised grammar induction using training data with limited constituent information. *CoRR*, cs.CL/9905001. Version 1.

Rebecca Hwa. 1999b. Supervised grammar induction using training data with limited constituent information. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, second edition. Pearson Prentice Hall.

Maxwell Mirton Kessler. 1963. Bibliographic coupling between scientific papers. *American documentation*, 14(1):10–25.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David Graham, Fangzhou Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler Murray, Chris Newell, Smita Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, Amber Tanaka, Alex D. Wade, Linda Wagner, Lucy Lu Wang, Chris Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine Van Zuylen, and Daniel S. Weld. 2023. The Semantic Scholar Open Data Platform. *arXiv e-prints*, page arXiv:2301.10140.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019a. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019b. Latent retrieval for weakly supervised open domain question answering. pages 6086–6096.

Adam Lerer, Ledell Wu, Jiajun Shen, Timothee Lacroix, Luca Wehrstedt, Abhijit Bose, and Alex Peysakhovich. 2019. PyTorch-BigGraph: A Large-scale Graph Embedding System. *arXiv e-prints*, page arXiv:1903.12287.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Weishu Liu. 2017. The changing role of non-english papers in scholarly communication: Evidence from web of science's three journal citation indexes. *Learned Publishing*, 30(2):115–123.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Melissa C Márquez and Ana Maria Porras. 2020. Science communication in multiple languages is critical to its effectiveness. *Frontiers in Communication*, page 31.

Aditya Menon, Sadeep Jayasumana, Ankit Singh Rawat, Seungyeon Kim, Sashank Reddi, and Sanjiv Kumar. 2022. In defense of dual-encoders for neural ranking. In *International Conference on Machine Learning*, pages 15376–15400. PMLR.

Olga Moskaleva and Mark Akoev. 2019. Non-english language publications in citation indexes–quantity and quality. In *17th International Conference on Scientometrics and Informetrics, ISSI 2019*, pages 35–46. International Society for Scientometrics and Informetrics.

Sheshera Mysore, Arman Cohan, and Tom Hope. 2022. Multi-vector models with textual guidance for fine-grained scientific document similarity. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4453–4470, Seattle, United States. Association for Computational Linguistics.

Jeppe Nicolaisen. 2007. Citation analysis. *Annual review of information science and technology*, 41(1):609–641.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11670–11688, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Silvio Peroni and David Shotton. 2020. OpenCitations, an infrastructure organization for open scholarship. *Quantitative Science Studies*, 1(1):428–444.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710.

Ştefan Postăvaru, Anton Tsitsulin, Filipe Miguel Gonçalves de Almeida, Yingtao Tian, Silvio Lattanzi, and Bryan Perozzi. 2020. Instantembedding: Efficient local node representations. *arXiv preprint arXiv:2010.06992*.

Derek J de Solla Price. 1963. *Little science, big science*. Columbia University Press.

Dragomir R. Radev, Pradeep Muthukrishnan, and Vahed Qazvinian. 2009. The ACL Anthology network. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries (NLPIR4DL)*, pages 54–61, Suntec City, Singapore. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Elizabeth Salesky, Kareem Darwish, Mohamed Al-Badrashiny, Mona Diab, and Jan Niehues. 2023. Evaluating multilingual speech translation under realistic conditions with resegmentation and terminology. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 62–78, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Henry Small. 1973. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science*, 24(4):265–269.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29.

Emma Steigerwald, Valeria Ramírez-Castañeda, Débora YC Brandt, András Báldi, Julie Teresa Shapiro, Lynne Bowker, and Rebecca D Tarvin. 2022. Overcoming language barriers in academia: machine translation tools and a vision for a multilingual future. *BioScience*, 72(10):988–998.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41, virtual+Dublin. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Sedef Uzuner. 2008. Multilingual scholars' participation in core/global academic communities: A literature review. *Journal of English for academic Purposes*, 7(4):250–263.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Amir Pouran Ben Veyseh, Nicole Meister, Seunghyun Yoon, Rajiv Jain, Franck Dernoncourt, and Thien Huu Nguyen. 2022. MACRONYM: A large-scale dataset for multilingual and multi-domain acronym extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3309–3314, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 363–372.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022a. A survey on cross-lingual summarization. *Transactions of the Association for Computational Linguistics*, 10:1304–1323.

Yau-Shian Wang, Ashley Wu, and Graham Neubig. 2022b. English contrastive learning can learn universal cross-lingual sentence embeddings. *arXiv preprint arXiv:2211.06127*.

Chih-Ping Wei, Christopher C Yang, and Chia-Min Lin. 2008. A latent semantic indexing-based approach to multilingual document clustering. *Decision Support Systems*, 45(3):606–620.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Xiang Wu, Ruiqi Guo, David Simcha, Dave Dopson, and Sanjiv Kumar. 2019. Efficient inner product approximation in hybrid spaces. *arXiv preprint arXiv:1903.08690*.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking Large Language Models for News Summarization. *arXiv e-prints*, page arXiv:2301.13848.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. NCLS: Neural cross-lingual summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3054–3064, Hong Kong, China. Association for Computational Linguistics.

Pranas Zunde. 1971. Structural models of complex information sources. *Information storage and retrieval*, 7(1):1–18.