S2ORC: The Semantic Scholar Open Research Corpus

Kyle Lo[†*] **Lucy Lu Wang**[†*] **Mark Neumann**[†] **Rodney Kinney**[†] **Daniel S. Weld**[†‡]
[†]Allen Institute for Artificial Intelligence
[‡]Paul G. Allen School of Computer Science & Engineering, University of Washington
{kylel, lucyw}@allenai.org

Abstract

We introduce S2ORC,[1] a large corpus of 81.1M English-language academic papers spanning many academic disciplines. The corpus consists of rich metadata, paper abstracts, resolved bibliographic references, as well as structured full text for 8.1M open access papers. Full text is annotated with automatically-detected inline mentions of citations, figures, and tables, each linked to their corresponding paper objects. In S2ORC, we aggregate papers from hundreds of academic publishers and digital archives into a unified source, and create the largest publicly-available collection of machine-readable academic text to date. We hope this resource will facilitate research and development of tools and tasks for text mining over academic text.

1 Introduction

Academic papers are an increasingly important textual domain for natural language processing (NLP) research. Aside from capturing valuable knowledge from humankind's collective research efforts, academic papers exhibit many interesting characteristics – thousands of words organized into sections, objects such as tables, figures and equations, frequent inline references to these objects, footnotes, other papers, and more.

Different types of resources have been used to support research over academic papers. Citation graphs like AMiner's Open Academic Graph (Tang et al., 2008), the Microsoft Academic Graph (MAG) (Shen et al., 2018), and the Semantic Scholar literature graph (Ammar et al., 2018), have had widespread application in bibliometrics, science-of-science, information retrieval, and network analysis. Digital archives like arXiv,[2]

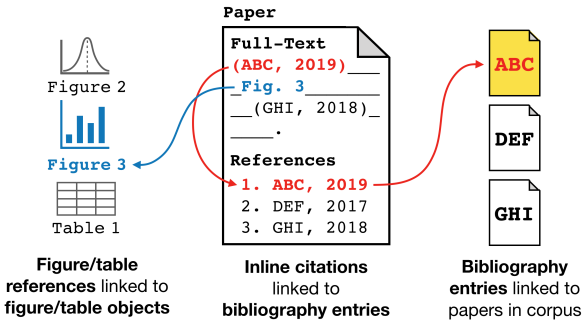

Figure 1: Inline citations and references to figures and tables are annotated in S2ORC's structured full text. Citations are linked to bibliography entries, which are linked to other papers in S2ORC. Figure and table references are linked to their captions.

PubMed Central,[3] CiteSeerX (Giles et al., 1998),[4] and the ACL Anthology (Bird et al., 2008),[5] are popular resources for deriving large text corpora for summarization and language modeling or, with further annotation, development of datasets for tasks like entity extraction, text classification, parsing, and discourse analysis. We focus on bibliometrically-enhanced derivations of these corpora, such as the ACL Anthology Network (AAN) (Radev et al., 2009)[6] derived from the ACL Anthology, RefSeer (Huang et al., 2015) derived from CiteSeerX, and Saier and Färber (2019) derived from arXiv, which combine useful aspects of citation graphs and raw text corpora. These resources provide citation mentions linked to paper identifiers in their corresponding digital archives, such as the ACL Anthology and CiteSeerX, or to nodes in citation graphs such as MAG, enabling new forms of cross-paper discourse analysis (e.g., studying *how* or *why* papers are related).

[*]denotes equal contribution
[1]Instructions for access to the data and model are available at https://github.com/allenai/s2orc/.
[2]https://arxiv.org
[3]https://www.ncbi.nlm.nih.gov/pmc
[4]https://citeseerx.ist.psu.edu
[5]https://www.aclweb.org/anthology
[6]http://aan.how/

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4969–4983
July 5 - 10, 2020.

Corpus	Papers w/ body text	Citation contexts	References to tables / figures / equations	Linked to graph	Academic disciplines
S2ORC (PDF-parse)	**8.1M**	**full text**	**yes**	**S2ORC (full)**	**multi**
S2ORC (LaTeX-parse)	**1.5M**	**full text**	**yes**	**S2ORC (full)**	**physics, math, CS**
PubMed Central (OA)	2.6M	full text	yes	PubMed	bio, med
AAN (Radev et al., 2009)	25k	full text	no	ACL Anthology	comp ling
Saier and Färber (2019)†	1.0M	snippets	no	MAG	physics, math, CS
RefSeer (Huang et al., 2015)	1.0M	snippets	no	CiteSeerX	multi

Table 1: A comparison of S2ORC with other publicly-available academic text corpora. Of the other corpora: PubMed Central (OA) links to PubMed, which contains 30M papers at the time of writing. AAN links to the ACL Anthology (which contained 25k papers at the time of dataset construction, and 54k papers at the time of writing). Saier and Färber (2019) is derived from arXiv and links to MAG (which contained 213M papers and other non-paper documents at the time of dataset construction, and 226M nodes at the time of writing). RefSeer links to CiteSeerX (which contained 1M papers at the time of dataset construction, and 6M papers at the time of writing). S2ORC contains three times more full text papers than PubMed Central (OA), the next largest corpus with bibliometric enhancements, while covering a more diverse set of academic disciplines. Citations in S2ORC are linked to the full set of S2ORC papers, 81.1M paper nodes derived from Semantic Scholar. In addition, the LaTeX subset of S2ORC captures additional structure omitted by Saier and Färber (2019), who also parse LaTeX sources from arXiv.

†Saier and Färber (2020) is an update to this work which now includes full text. It is released concurrently with this work.

Yet, existing corpora are not without their limitations. Some cover a small number of papers (e.g. AAN), are domain-specific (e.g. AAN, PubMed Central, Saier and Färber (2019)), or may not provide usable full text (e.g. Saier and Färber (2019) and RefSeer). To address these issues, we introduce S2ORC,[7] the Semantic Scholar[8] Open Research Corpus, a large publicly-available collection of 81.1M academic papers covering dozens of academic disciplines. Each paper is associated with metadata and abstracts aggregated from hundreds of trusted sources such as academic publishers and literature archives like PubMed and arXiv.

Notably, we release structured, machine-readable full text extracted from PDFs for 8.1M papers which we've identified as having open access status. S2ORC full text preserves meaningful structure, e.g., paragraph breaks, section headers, inline citation mentions, references to tables and figures, and resolved citation links to other papers. Additionally, we provide 1.5M full text LaTeX parses from which we have extracted, in addition to citations and references, the source text of tables and mathematical formulas. As shown in Table 1, S2ORC provides substantially more structured full text papers and covers a more diverse set of academic disciplines than other resources.

[7]pronounced "stork"

[8]The papers included in S2ORC are a curated subset of the papers in the Semantic Scholar literature graph (Ammar et al., 2018) that focuses only on English-language papers with abstracts or full text available. See §2.5 for details on filtering through Semantic Scholar papers.

In this paper, we describe the construction of S2ORC (§2). We provide summary statistics of the corpus (§3) and evaluate the data quality (§4). We then evaluate a BERT model pretrained on S2ORC (§5), and discuss potential applications to a variety of NLP and analysis tasks over academic text (§6). Finally, we compare S2ORC with other publicly-available academic text corpora (§7).

2 Constructing the corpus

S2ORC is constructed using data from the Semantic Scholar literature corpus (Ammar et al., 2018). Papers in Semantic Scholar are derived from numerous sources: obtained directly from publishers, from resources such as MAG, from various archives such as arXiv or PubMed, or crawled from the open Internet. Semantic Scholar clusters these papers based on title similarity and DOI overlap, resulting in an initial set of approximately 200M paper clusters.

To construct S2ORC, we must overcome challenges in (i) paper metadata aggregation, (ii) identifying open access publications, and (iii) clustering papers, in addition to identifying, extracting, and cleaning the full text and bibliometric annotations associated with each paper. The pipeline for creating S2ORC is:

1) Process PDFs and LaTeX sources to derive metadata, clean full text, inline citations and references, and bibliography entries,
2) Select the best metadata and full text parses for each paper cluster,

3) Filter paper clusters with insufficient metadata or content, and
4) Resolve bibliography links between paper clusters in the corpus.

Details for these steps are provided below. See Appendix §A for definitions of terminology. The output of this pipeline is visualized in Figure 1.

2.1 Processing PDFs

We process PDFs from the Semantic Scholar corpus using SCIENCEPARSE v3.0.0[9] and GROBID v0.5.5[10] (Lopez, 2009). Our processing pipeline is described below.

Selecting PDFs We remove PDFs which are less likely to be academic papers. SCIENCEPARSE and GROBID are not optimized for processing non-paper academic documents such as dissertations, reports, slides, etc., and this filtering step is necessary to increase output data quality. See Appendix §B for filter details. There are around 31.3M PDFs associated with approximately 200M initial paper clusters, and 30.5M PDFs are selected for processing based on these filtering criteria.

Extracting structured data from PDFs We use SCIENCEPARSE to extract title and authors from each PDF.[11] We then use GROBID to process each PDF. From the XML output of GROBID, we extract (i) metadata such as title, authors, and abstract, (ii) paragraphs from the body text organized under section headings, (iii) figure and table captions, (iv) equations, table content, headers, and footers, which we remove from the body text, (v) inline citations in the abstract and body text, (vi) parsed bibliography entries with title, authors, year, and venue identified, and (vi) links between inline citation mentions and their corresponding bibliography entries.

Postprocessing GROBID output We postprocess GROBID output using regular expressions to classify the parenthetical citation style of a paper as BRACKET (e.g. [2]), NAME-YEAR (e.g. ABC, 2019), or OTHER (superscripts and other mixed styles). We focus on addressing two types of common errors in GROBID's inline citation extractions: (i) false positives resulting from superscripts or equation references being recognized as inline citations in papers with BRACKET-style citations, and (ii) false negatives resulting from an inability to expand bracket citation ranges (e.g. "[3]-[5]" should be expanded to "[3], [4], [5]" before linking). False positives are detected using regular expressions and removed from GROBID output. Bracket citation ranges are manually expanded and linked to their corresponding bibliography entries. The resulting parses are expressed in JSON format.[12]

2.2 Processing LATEX source

LATEX document source is available for a majority of arXiv submissions, and where available, are used to construct a full text parse. We retrieve body text, section headers, figure/table captions, table representations, equations, and inline citations and references directly from LATEX source. Inspired by Saier and Färber (2019), we first convert LATEX source into XML documents and then extract structured information from the XML.

Due to direct access to source, the accuracy of citation span, reference, caption, section header, and equation detection is near-perfect. We process 1.5M papers from LATEX source derived from arXiv, all of which are included as part of S2ORC. Surprisingly, due to the diversity of ways in which authors define metadata in LATEX, the quality of metadata extracted from LATEX documents is worse than those extracted from PDF. Therefore, we do not use LATEX-derived metadata for paper clustering or metadata selection.

2.3 Selecting canonical metadata

Canonical values for title, authors and other metadata fields are selected from among the papers in a cluster. First, if a cluster contains multiple PDFs, we select one to be canonical. This can occur, for example, in a cluster containing an arXiv preprint and its eventual camera-ready version. We preferentially select PDFs from open access sources and break ties by prioritizing PDFs for which there exist richer publisher-provided metadata (e.g. abstract, year, venue, DOI). If the selected PDF is associated with publisher-provided metadata, we select those publisher-provided metadata fields to be canonical.

In cases where publisher-provided metadata is incomplete, we use majority voting to select

[9] https://github.com/allenai/science-parse

[10] https://github.com/kermitt2/grobid

[11] Our evaluations suggest SCIENCEPARSE outperforms GROBID for title and author extraction.

[12] The S2ORC data format is described at https://github.com/allenai/s2orc

canonical metadata values. We break ties by minimizing the total number of sources from which we select metadata (e.g., if IEEE provides title, authors and abstract, DBLP provides title and authors, and arXiv provides title and abstract, we prioritize selecting IEEE over the union of DBLP and arXiv). S2ORC metadata fields include title, author, year, venue, journal, abstract, and identifiers (DOI, PubMed, PubMed Central (PMC), arXiv, and ACL Anthology).

In cases where the title and authors are not provided by any publishers, we derive the values for these fields from the parsed PDF, prioritizing SCIENCEPARSE over GROBID. We further comment on paper clustering as it pertains to metadata selection in Appendix §C.

2.4 Assembling the corpus

We construct the final corpus by assembling clustered paper metadata with GROBID and LATEX parse objects. We associate the GROBID parse with the S2ORC paper object if a valid GROBID parse is produced from the PDF, and the PDF is open access. Open access status is assigned if a paper is derived from arXiv, ACL Anthology, PubMed Central (OA), and/or associated with an open-access DOI in the Unpaywall database.[13] If the PDF is not open access, we only include the bibliography from the GROBID parse in S2ORC. If arXiv LATEX source is available for the paper cluster, we also associate the LATEX parse with the S2ORC paper object.

2.5 Filtering paper clusters

We further filter paper clusters to remove papers with (i) no title, (ii) no authors, (iii) fewer than 100 characters of abstract and body text, and (iv) where English is not the primary language. The first three filters remove papers that provide little value for bibliometric-based or text-based analyses. The English language filter[14] reduces GROBID parsing errors. All filters are applied in series.

Subsequently, 95.5M paper clusters are filtered out based on the aforementioned criteria and removed from the corpus. The distribution of filtered papers is given in Table 2. We note that a large number of paper clusters are filtered out; 80.0M of these filtered clusters have no associated publisher-provided abstract or associated PDF and do not provide significant value to our dataset in their current state. Although these papers that lack text may be useful as cite-able nodes in S2ORC, they are generally of lower quality and are filtered out of the corpus to improve corpus quality.

Filter	Number of papers
No title	20k
No authors	0.3M
$<$ 100 chars of text	80.0M
Not English	15.2M

Table 2: Post-processing data quality filters for papers

2.6 Linking bibliographies to papers

Each bibliography entry in both GROBID and LATEX parses are linked to the most similar papers in the corpus. For linking, we score each bibliography entry and paper cluster pair using a similarity score computed between their titles. Each title is first normalized (i.e. white spaces stripped, lower-cased, special characters removed) and represented by its character 3-grams. The similarity score S_{title} is computed as the harmonic mean between a Jaccard index and a containment metric:

$$S_{title} = \frac{2 \times J \times C}{J + C} \quad (1)$$

where the Jaccard index J and containment metric C are computed from the n-grams of the two titles N_1 and N_2 as:

$$J = \frac{|N_1 \cap N_2|}{|N_1 \cup N_2|}$$

$$C = \frac{|N_1 \cap N_2|}{\min(|N_1|, |N_2|)}$$

For each bibliography entry, the bibliography-paper pair with the highest similarity score above 0.8 is output as the correct link. Otherwise, the bibliography entry remains unlinked. We perform an evaluation of linking performance in §4.

3 The S2ORC dataset

The resulting corpus consists of 81.1M papers. Our publisher-provided abstract coverage is 90.4%, or 73.4M papers. Our PDF coverage is 35.6%, or 28.9M papers. These PDFs are processed using the pipeline discussed in §2.1. The

[13] Unpaywall 2019-04-19 data dump

[14] We use the `cld2` tool for language detection with a threshold of 0.9 over the English language score.

Total papers	81.1M
Papers w/ PDF	28.9M (35.6%)
Papers w/ bibliographies	27.6M (34.1%)
Papers w/ GROBID full text	8.1M (10.0%)
Papers w/ LaTeX full text	1.5M (1.8%)
Papers w/ publisher abstract	73.4M (90.4%)
Papers w/ DOIs	52.2M (64.3%)
Papers w/ Pubmed IDs	21.5M (26.5%)
Papers w/ PMC IDs	4.7M (5.8%)
Papers w/ ArXiv IDs	1.7M (2.0%)
Papers w/ ACL IDs	42k (0.1%)

Table 3: Statistics on paper provenance. We note that categories are not mutually exclusive and do not sum to 100%. All papers in S2ORC have either a publisher-provided abstract or an associated PDF from which we derive full text and/or bibliography entries, or both.

Statistic	GROBID	LATEX
Paragraphs (abstract)	1.1	-
Paragraphs (body)	9.9	93.3*
Inline cite spans (abstract)	0.7	-
Inline cite spans (body)	45.2	46.8
Bibliography entries	27.6	21.9
Linked bib. entries	19.3	6.8†

Table 4: Extraction and linking statistics over PDF and LATEX parses. Reported values are averaged over all open access papers, which consist of 8.1M GROBID-parsed PDFs and 1.5M parsed LATEX sources.

*LATEX preserves line breaks rather than paragraph breaks.

†The lower number of linked bibliography entries in LATEX parses is due to large numbers of papers (mostly in the field of physics) for which the bibliography entries are formatted without paper titles. Our linking algorithm strongly depends on titles and fails to link these entries.

vast majority of these PDFs are successfully processed using GROBID, and we extract bibliography entries for 27.6M of the 28.9M PDFs. We identify 8.1M of the 28.9M PDFs as open access (§2.4), and we provide full text for all papers in this open access subset. For the 1.5M papers for which LATEX source is available through arXiv, we further obtain and provide LATEX parses (§2.2). Using these extracted bibliographies, we resolve a total 380.5M citation links between papers (§2.6), 156.5M of which can be tied back to their inline citation mentions in the full text. See Table 3 for more provenance statistics.

We provide statistics for the GROBID and LATEX full text parses and bibliography linking in Table 4. On average, LATEX parses contain many more "paragraphs" of body text, because LATEX source files preserve line breaks rather than paragraph breaks. We speculate that differences in bibliography entry and linking counts between the GROBID and LATEX parses are due to a combination of: (i) challenges in LATEX bibliography expansion and parsing, and (ii) differences in bibliography formatting in some math and physics venues (where bibliography entries do not include paper titles, which we depend on for bibliography linking).

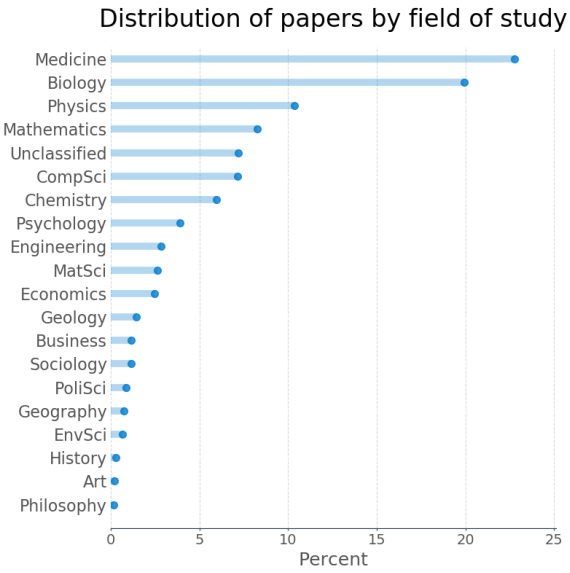

Figure 2: Distribution of papers by Microsoft Academic field of study.

The distribution of academic disciplines in S2ORC is given in Figure 2 using Microsoft Academic fields of study. Not all papers in S2ORC can be found in Microsoft Academic – those not found are denoted as *Unclassified*. Approximately 677k papers have more than one primary Microsoft Academic field of study; Figure 2 represents only the top field of study for each paper.

4 Evaluation

To evaluate the quality of our metadata selection, we randomly sample 500 paper clusters, restricting to those with PDFs. Within each sampled cluster, we determine whether the canonical title and authors match the title and authors in the selected canonical PDF.

Inline citation detection and bibliography parsing are dependent on GROBID (Lopez, 2009). Ahmad and Afzal (2018) evaluate GROBID for de-

Domain	Dataset	Reference	Task	SciBERT	S2ORC-SciBERT
Biomed	BC5CDR	Li et al. (2016)	NER	90.01	90.41 ± 0.06
	JNLPBA	Collier and Kim (2004)	NER	77.28	77.70 ± 0.25
	NCBI-disease	Doğan et al. (2014)	NER	88.57	88.70 ± 0.52
	EBM-NLP	Nye et al. (2018)	PICO	72.28	72.35 ± 0.95
	GENIA	Kim et al. (2003)	DEP (LAS)	90.43	90.80 ± 0.19
	GENIA	Kim et al. (2003)	DEP (UAS)	91.99	92.31 ± 0.18
	ChemProt	Krallinger et al. (2017)	REL	83.64	84.59 ± 0.93
CS	SciERC	Luan et al. (2018)	NER	67.57	68.93 ± 0.19
	SciERC	Luan et al. (2018)	REL	79.97	81.77 ± 1.64
	ACL-ARC	Jurgens et al. (2018)	CLS	70.98	68.45 ± 2.47
Biomed & CS	SciCite	Cohan et al. (2019)	CLS	85.49	84.76 ± 0.37
Multi-domain	PaperField	Beltagy et al. (2019)	CLS	65.71	65.99 ± 0.08

Table 5: S2ORC-SciBERT test results are comparable with reported SciBERT test results on the set of tasks and datasets from Beltagy et al. (2019), to which we refer the reader for descriptions. Reported statistics are span-level F1 for NER, token-level F1 for PICO, dependency parsing (DEP), and macro-F1 for relation (REL) and text (CLS) classification. We report micro-F1 for ChemProt. All S2ORC-SciBERT results are the mean ± standard deviation of 5 runs with different random seeds. Beltagy et al. (2019) do not report standard deviation or number of runs.

tecting inline citations using a corpus of 5k CiteSeer papers, and found GROBID to have an F1-score of 0.89 on this task. Tkaczyk et al. (2018) report GROBID as the best among 10 out-of-the-box tools for parsing bibliographies, also achieving an F1 of 0.89 in an evaluation corpus of 9.5k papers. We perform an evaluation over 200 randomly sampled papers from S2ORC and found comparable F1-scores for GROBID performance on both tasks.

For bibliography linking, we randomly sample S2ORC papers (500 GROBID PDF parses and 100 LATEX parses) and select one linked bibliography entry from each sampled paper (while avoiding selecting multiple entries linked to the same paper). We determine whether the title and authors in the bibliography entry agree with the title and authors of the linked paper.

We present these evaluation results in Table 6 and detail valuation criteria in Appendix §D.

Evaluated task	Title	Authors
Paper clustering	0.93	0.89
Bib. linking (GROBID)	1.00	0.96
Bib. linking (LATEX)	1.00	0.92

Table 6: Accuracy of paper clustering and bibliography linking for titles and authors in sampled evaluation sets.

5 Pretraining BERT on S2ORC

To demonstrate the suitability of S2ORC for language model pretraining, we train BERT-Base (Devlin et al., 2019) on the parsed full text of S2ORC and show that the resulting model (S2ORC-SciBERT) performs similarly to SciBERT (Beltagy et al., 2019) on a diverse suite of scientific NLP tasks and datasets.

While SciBERT is a BERT-Base model also trained on multiple domains of scientific text, key differences in its pretraining corpus and vocabulary and those used for S2ORC-SciBERT are:

- **Domain:** Beltagy et al. (2019) report a pretraining corpus consisting of 82% biomedical and 18% computer science papers. Our S2ORC pretraining corpus consists of a more balanced distribution of papers across diverse academic disciplines (see Figure 2), such that biomedical (42.7%) and computer science (7.2%) papers only comprise half the corpus.
- **Preprocessing:** S2ORC identifies figure captions, table text and captions, headers, footers, and footnotes. We exclude these from the pretraining corpus. We tokenize and sentencize the text using scispaCy (Neumann et al., 2019). We also use heuristic filters to remove ill-formed paragraphs (such as those containing too many symbols).
- **Size:** The resulting S2ORC pretraining cor-

pus contains 16.4B tokens, nearly five times larger than the corpus for SciBERT.

- **Vocab:** Following Beltagy et al. (2019), we construct a cased WordPiece (Wu et al., 2016) vocabulary of size 31k using 15% of the S2ORC pretraining corpus. The Jaccard index between the S2ORC-SciBERT and SciBERT vocabularies is 0.536.

We follow a similar setup to Beltagy et al. (2019) for both pretraining and fine-tuning S2ORC-SciBERT. Like SciBERT, S2ORC-SciBERT is pretrained from scratch using the original BERT code[15] and default BERT-Base configurations on a single TPU v3-8 for one week. Also like SciBERT, S2ORC-SciBERT is fine-tuned on all tasks by optimizing a cross entropy loss using Adam (Kingma and Ba, 2014), a linear learning rate decay with 10% warm-up, batch size of 32, and dropout of 0.1.

We search over an equal-sized grid of hyperparameters as Beltagy et al. (2019). We fine-tune for 1 to 4 epochs with a maximum learning rate of 1e-5, 2e-5, 3e-5, or 5e-5. For each task, we select the optimal combination of these two hyperparameters using the development set and report the corresponding test set results. For details, we refer the reader to SciBERT code,[16] which we use for all experiments.

The results in Table 5 show that S2ORC-SciBERT outperforms SciBERT on many tasks despite including a large percentage of data outside of the biomedical and computer science domains. As the pretraining corpus for SciBERT is not publicly-available, S2ORC can serve as a large pretraining corpus for evaluating and comparing pretraining approaches on academic text. We also release S2ORC-SciBERT to serve as a baseline for research.

6 Applications of S2ORC

S2ORC can be used for many NLP and analysis tasks over academic text. We give a summary of potential applications below.

The combination of structured full text annotated with linked inline citations makes S2ORC well-suited for a variety of citation-related text-based tasks. Without any additional supervision, S2ORC can be used directly for both inline (He et al., 2010; Duma and Klein, 2014; Jeong et al., 2019) and document-level (Yu et al., 2012; Liu et al., 2015; Bhagavatula et al., 2018) citation recommendation. Among document-level recommenders, S2ORC is well-suited to the setting of Liu et al. (2015), who use inline citation contexts to filter document-level recommendations.

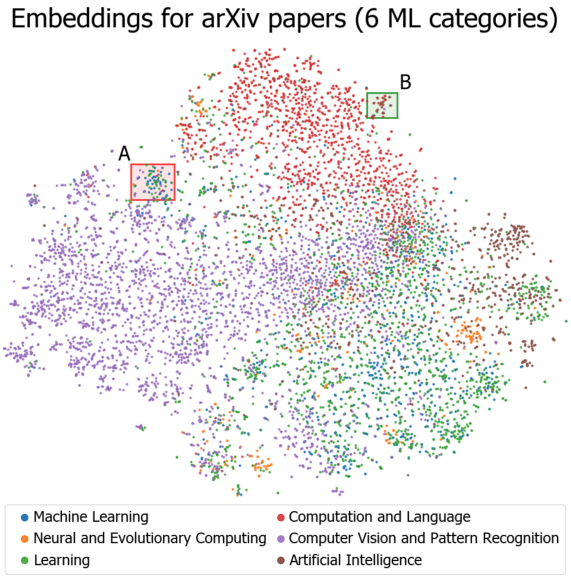

Figure 3: *Word2vec* embeddings associated with 20k papers in six AI-related arXiv categories visualized using t-SNE (van der Maaten and Hinton, 2008). Example papers from two randomly selected sub-regions A and B are given in Table 7.

Region A	
cs.LG	"On Unifying Deep Generative Models"
stat.ML	"Learning Disentangled Representations with Semi-Supervised Deep Generative Models"
cs.LG	"Denoising Criterion for Variational Auto-Encoding Framework"
cs.CV	"Variational methods for conditional multimodal deep learning"
Region B	
cs.CL	"TransA: An Adaptive Approach for Knowledge Graph Embedding"
cs.AI	"TorusE: Knowledge Graph Embedding on a Lie Group"
cs.CV	"Image-embodied Knowledge Representation Learning"
stat.ML	"Neural Embeddings of Graphs in Hyperbolic Space"

Table 7: Sampled papers in clusters from t-SNE embedding space in Figure 3. Region A consists of papers related to deep generative models; region B consists of papers concerned with graph representation learning.

Other tasks that leverage citation contexts in-

[15] https://github.com/google-research/bert

[16] https://github.com/allenai/scibert

clude classifying citation intent (Teufel et al., 2006; Jurgens et al., 2018; Cohan et al., 2019), identifying citation sentiment (Athar and Teufel, 2012), identifying meaningful citations (Valenzuela et al., 2015), extracting key phrases (Caragea et al., 2014), and citation context-based paper summarization (Teufel et al., 2006; Qazvinian and Radev, 2008; Cohan and Goharian, 2015; Mitrović and Müller, 2015). The models in these papers require labeled citation contexts for training. S2ORC could potentially benefit task performance without additional annotation, for example, by pretraining language models on S2ORC citation contexts before fine-tuning to these tasks. Cohan et al. (2019) find that long citation contexts (beyond sentence boundary) are important for tasks like summarization; the wider citation contexts available in S2ORC could be used to augment existing datasets for document-level tasks.

Citation contexts can also be used for the more general tasks of identifying similar papers (Kanakia et al., 2019; Eto, 2019; Haruna et al., 2018; Small, 1973) or bibliometric analysis (Ding et al., 2014; Trujillo and Long, 2018; Asatani et al., 2018). Towards these tasks, the citation contexts in S2ORC can provide insight into *how* and *why* papers are cited. We illustrate this by following Berger et al. (2016) in training a *word2vec* skip-gram model (Mikolov et al., 2013) using full text citation contexts in S2ORC, where each inline citation span is replaced with its linked paper identifier. When training over this modified text, the *word2vec* model learns embeddings corresponding to each unique paper identifier, which can be leveraged as paper embeddings. The resulting embeddings shown in Figure 3 and Table 7 form clusters corresponding closely to arXiv Machine Learning categories. Upon inspection, papers of different categories in the same embedding sub-region share research themes (see Table 7), indicating that these paper embeddings trained from citation contexts capture coherent topic similarity and relatedness. These paper embeddings can be used to identify similar papers, using the similarity between two papers' citing contexts as a proxy for paper similarity.

The LaTeX subset of S2ORC also provides unique opportunities for research. In addition to citations and references, we also extract and parse tables from LaTeX source into a structured format. There is an opportunity to use these tables for corpus-level results extraction and aggregation. The LaTeX subset also has fine-grained extraction and labeling of mathematical formulas, which can be used to understand proof construction, or to assist in symbol co-reference resolution.

7 Related work

The ACL Anthology Network (AAN) (Radev et al., 2009) is a bibliometric-enhanced corpus covering papers in the field of computational linguistics. It is built from the ACL Anthology (Bird et al., 2008) and consists of 24.6k papers manually augmented with citation information. The PubMed Central Open Access corpus is a large corpus of 2.6M papers in the biomedical domain with citations linked to PubMed identifiers.[17] CiteSeerX (Giles et al., 1998), consists of papers collected primarily via web crawl, without integrating metadata provided by sources outside of the PDF. Although citation contexts are no longer available through CiteSeerX, the RefSeer dataset (Huang et al., 2015)[18] is a dataset of short citation context snippets derived from 1.0M papers from CiteSeerX. More recently, Saier and Färber (2019) introduce a corpus built using 1.0M arXiv publications. They use LaTeX source to extract text, citation spans and bibliography entries, which are linked to papers in the Microsoft Academic Graph. The citation context they provide are extracted snippets and no bibliography parses are provided. An updated version of this dataset (Saier and Färber, 2020) released concurrently with this work now includes full text.

Compared with these resources, S2ORC represents a significantly larger dataset of linked papers covering broad domains of science by leveraging PDF parsing in addition to LaTeX source. S2ORC also provides clean full text for text mining and NLP needs with additional enhancements such as annotations of table and figure references and captions. S2ORC's wealth of metadata and structured text allows it to be flexibly adapted to a variety of downstream tasks.

8 Conclusion

We introduce S2ORC, the largest publicly-available corpus of English-language academic papers covering dozens of academic disciplines.

[17] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

[18] https://psu.app.box.com/v/refseer

S2ORC consists of 81.1M papers, 380.5M resolved citation links, and structured full text from 8.1M open-access PDFs and 1.5M LaTeX source files. We aggregate metadata and abstracts from hundreds of trusted sources. Full text is augmented with sections, citation mentions, and references to tables and figures. We demonstrate that S2ORC can be used effectively for downstream NLP tasks in academic paper analysis.

The pipeline for creating S2ORC was used to construct the CORD-19 corpus (Wang et al., 2020), which saw fervent adoption as the canonical resource for COVID-19 text mining. CORD-19 is aimed at assisting biomedical experts and policy makers process large amounts of COVID-19 literature in the search for effective treatments and management policies. With over 75K dataset downloads, dozens of search and question-answering systems, and hundreds of participating teams across two shared tasks[19] in the first month of its release, there is little doubt of the resource's impact. Our hope with the release of S2ORC is to ensure such text mining resources are available to researchers even beyond periods of global crisis.

Acknowledgements

This work was supported in part by ONR grant N00014-18-1-2193, and the University of Washington WRF/Cable Professorship.

We thank Doug Downey, Oren Etzioni, Andrew Head, and Bryan Newbold for their valuable feedback on the manuscript. We also thank Isabel Cachola, Dallas Card, Mike D'Arcy, Suchin Gururangan, Daniel King, Rik Koncel-Kedziorski, Susan Liu, Kelvin Luu, Noah Smith, Gabi Stanovsky, and Dave Wadden for feedback on the dataset during early development. Finally, we thank the Semantic Scholar team for assisting with data access and system infrastructure.

References

Riaz Ahmad and Muhammad Tanvir Afzal. 2018. Cad: an algorithm for citation-anchors detection in research papers. *Scientometrics*, 117:1405–1423.

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. 2018. Construction of the literature graph in semantic scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 84–91, New Orleans - Louisiana. Association for Computational Linguistics.

Kimitaka Asatani, Junichiro Mori, Masanao Ochi, and Ichiro Sakata. 2018. Detecting trends in academic research from a citation network using network representation learning. In *PloS one*.

Awais Athar and Simone Teufel. 2012. Context-enhanced citation sentiment detection. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 597–601, Montréal, Canada. Association for Computational Linguistics.

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.

Matthew Berger, Katherine McDonough, and Lee M Seversky. 2016. cite2vec: Citation-driven document exploration via word embeddings. *IEEE transactions on visualization and computer graphics*, 23(1):691–700.

Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. 2018. Content-based citation recommendation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 238–251, New Orleans, Louisiana. Association for Computational Linguistics.

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Cornelia Caragea, Florin Adrian Bulgarov, Andreea Godea, and Sujatha Das Gollapalli. 2014. Citation-enhanced keyphrase extraction from research papers: A supervised approach. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1435–1446, Doha, Qatar. Association for Computational Linguistics.

[19] The Kaggle CORD-19 and TREC-COVID competitions. See Wang et al. (2020) for details.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. 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 3586–3596, Minneapolis, Minnesota. Association for Computational Linguistics.

Arman Cohan and Nazli Goharian. 2015. Scientific article summarization using citation-context and article's discourse structure. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 390–400, Lisbon, Portugal. Association for Computational Linguistics.

Nigel Collier and Jin-Dong Kim. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 73–78, Geneva, Switzerland. COLING.

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

Ying Ding, Guo Zhang, Tamy Chambers, Min Song, Xiaolong Wang, and Cheng xiang Zhai. 2014. Content-based citation analysis: The next generation of citation analysis. *JASIST*, 65:1820–1833.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Daniel Duma and Ewan Klein. 2014. Citation resolution: A method for evaluating context-based citation recommendation systems. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 358–363, Baltimore, Maryland. Association for Computational Linguistics.

Masaki Eto. 2019. Extended co-citation search: Graph-based document retrieval on a co-citation network containing citation context information. *Inf. Process. Manage.*, 56.

C. L. Giles, K. D. Bollacker, and S. Lawrence. 1998. Citeseer: an automatic citation indexing system. In *Proceedings of the ACM International Conference on Digital Libraries*, pages 89–98. ACM. Proceedings of the 1998 3rd ACM Conference on Digital Libraries ; Conference date: 23-06-1998 Through 26-06-1998.

Khalid Haruna, Maizatul Akmar Ismail, Abdullahi Baffa Bichi, Victor I. Chang, Sutrisna Wibawa, and Tutut Herawan. 2018. A citation-based recommender system for scholarly paper recommendation. In *ICCSA*.

Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. 2010. Context-aware citation recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 421–430, New York, NY, USA. Association for Computing Machinery.

Wenyi Huang, Zhaohui Wu, Chen Liang, Prasenjit Mitra, and C. Lee Giles. 2015. A neural probabilistic model for context based citation recommendation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2404–2410. AAAI Press.

Chanwoo Jeong, Sion Jang, Hyuna Shin, Eunjeong Park, and Sungchul Choi. 2019. A context-aware citation recommendation model with bert and graph convolutional networks. *arXiv*.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Anshul Kanakia, Zhihong Shen, Darrin Eide, and Kuansan Wang. 2019. A scalable hybrid research paper recommender system for microsoft academic. In *The World Wide Web Conference*, WWW '19, page 2893–2899, New York, NY, USA. Association for Computing Machinery.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Martin Krallinger, Obdulia Rabal, Saber Ahmad Akhondi, Martín Pérez Pérez, Jésús López Santamaría, Gael Pérez Rodríguez, Georgios Tsatsaronis, Ander Intxaurrondo, José Antonio Baso López, Umesh Nandal, Erin M. van Buel, A. Poorna Chandrasekhar, Marleen Rodenburg, Astrid Lægreid, Marius A. Doornenbal, Julen Oyarzábal, Anália Lourenço, and Alfonso Valencia. 2017. Overview of the biocreative vi chemical-protein interaction track. In *N/A*.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Haifeng Liu, Xiangjie Kong, Xiaomei Bai, Wei Wang, Teshome Megersa Bekele, and Feng Xia. 2015. Context-based collaborative filtering for citation recommendation. *IEEE Access*, 3:1695–1703.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. 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 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke S. Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*.

Patrice Lopez. 2009. Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL'09, page 473–474, Berlin, Heidelberg. Springer-Verlag.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey E. Hinton. 2008. Visualizing data using t-sne. In *Journal of Machine Learning Research*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Sandra Mitrović and Henning Müller. 2015. Summarizing citation contexts of scientific publications. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 154–165, Cham. Springer International Publishing.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Melbourne, Australia. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018b. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Vahed Qazvinian and Dragomir R. Radev. 2008. Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 689–696, Manchester, UK. Coling 2008 Organizing Committee.

Dragomir R. Radev, Pradeep Muthukrishnan, and Vahed Qazvinian. 2009. The acl anthology network corpus. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, NLPIR4DL '09, page 54–61, USA. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *N/A*.

Tarek Saier and Michael Färber. 2019. Bibliometric-enhanced arxiv: A data set for paper-based and citation-based tasks. In *Proceedings of the 8th International Workshop on Bibliometric-enhanced Information Retrieval (BIR 2019) co-located with the 41st European Conference on Information Retrieval (ECIR 2019), Cologne, Germany, April 14, 2019*, volume 2345 of *CEUR Workshop Proceedings*, pages 14–26. CEUR-WS.org.

Tarek Saier and Michael Färber. 2020. unarxive: a large scholarly data set with publications' full-text, annotated in-text citations, and links to metadata. *Scientometrics*.

Zhihong Shen, Hao Ma, and Kuansan Wang. 2018. A web-scale system for scientific knowledge exploration. In *Proceedings of ACL 2018, System Demonstrations*, pages 87–92, Melbourne, Australia. Association for Computational Linguistics.

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.

Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnetminer: Extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 990–998, New York, NY, USA. Association for Computing Machinery.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 103–110, Sydney, Australia. Association for Computational Linguistics.

Dominika Tkaczyk, Andrew Collins, Paraic Sheridan, and Joeran Beel. 2018. Machine learning vs. rules and out-of-the-box vs. retrained: An evaluation of open-source bibliographic reference and citation parsers. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, JCDL '18, page 99–108, New York, NY, USA. Association for Computing Machinery.

Caleb M. Trujillo and Tammy M. Long. 2018. Document co-citation analysis to enhance transdisciplinary research. *Science Advances*, 4(1).

Marco Valenzuela, Vu Ha, and Oren Etzioni. 2015. Identifying meaningful citations. *AAAI*.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The Covid-19 Open Research Dataset.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv*, abs/1609.08144.

Xiao Yu, Quanquan Gu, Mianwei Zhou, and Jiawei Han. 2012. Citation prediction in heterogeneous bibliographic networks. In *SDM*.

A Background & Terminology

In this work, we distinguish between bibliography entries and inline citations. A **bibliography entry** is an item in a paper's bibliography that refers to another paper. It is represented in a structured format that can be used for paper-identifying features such as title, authors, year, and venue or journal, and for journal articles, the volume, issue, and pages. Also commonly represented are unique document identifiers such as the Document Object Identifier (DOI), arXiv identifier, or PubMed identifier. Common formats for bibliography entries are MLA, APA, Vancouver-, and Chicago- style, among others, which are different ways of representing these various features for document identification.

There is often variation in the representation of certain fields. For example, *Authors* can include the first names of each author or only their first initials. In many academic disciplines, journal publications are the norm, whereas conference proceedings dominate in fields such as Computer Science; conference proceedings tend to lack journal-related features such as *Volume*, *Issue*, and *Pages*. Bibliography entry demarcation also varies between different formats. In some cases, each entry is preceded by a citation marker (e.g. "[1]" or "[ABC2019]") that is used throughout the text of the paper to denote inline citations.

An **inline citation** is a mention span within the paper's abstract or body text that refers to one of the entries in its bibliography.

> "ABC (2019) present model 1, which outperforms model 2 (XYZ (2019))."

In this example, the **narrative** inline citation ABC (2019) appears as a noun phrase in the sentence while the **parenthetical** inline citation (XYZ, 2019) is inserted into the sentence as an aside. A sentence remains grammatically correct when parenthetical citations are removed. Other styles of parenthetical citations include, but are not limited to, BRACKET-style numbers (e.g. "[1, 3-5]") and OTHER styles such as superscripts (e.g. "[1,2]"), both of which refer to numbered entries in the bibliography. Bibliography entries without numbered entries or citation markers are typically referenced inline using NAME-YEAR format as ABC (2019) or (XYZ, 2019) in the example above.

Additionally, an **inline reference** is a span in a paper that refers to another part of the paper, for example, references to figures, tables, equations, proofs, sections, or appendices. These often take on the form of:

> "In Figure 3, we show the relationship between A and B."

where Figure 3 refers to a plot displayed on a separate page. These inline references can be important for understanding the relationship between text and objects within the paper.

B PDF filters

Prior to running GROBID, we filter out PDFs that (i) produce an error when processed using the Python library PyPDF2,[20] (ii) have greater than 50 pages (more likely to be a dissertation or report), (iii) have page widths greater than page heights (more likely to be slides), and (iv) those which fail to be extracted using pdfalto, the variant of pdftoxml used by GROBID.

Numbers of PDFs removed by these filters are given in Table 8.

Filter	Number of PDFs
PyPDF2 error	0.54M
Over 50 pages	2.27M
Page width > height	0.28M
PDFAlto error	0.21M

Table 8: PDFs filtered out before GROBID processing

C The paper clustering problem

In academic fields in which preprint publishing is common (e.g. arXiv), the notion of a "paper" is somewhat ambiguous. For example, if a published paper differs from its arXiv preprint (as it often does), are the two documents considered separate papers for the purposes of citation? What about different arXiv preprint drafts tagged as different versions but under the same arXiv identifier?

In this work, each "paper" of interest is actually a collection (or cluster) of highly-similar (but not necessarily identical) documents. These paper clusters, provided by Semantic Scholar, are constructed to reflect how authors tend to view their

[20] Used to determine PDF page number and page dimensions

own papers; for example, most authors would consider their arXiv preprint and its associated published version to be the same "paper". For practical concerns in constructing S2ORC, we further require that one document within the cluster be the canonical document used to represent the paper cluster.

There are issues with defining a paper to be a collection of documents. For example, suppose a paper cluster contains both an arXiv preprint and a peer-reviewed draft. And suppose another paper cites the arXiv preprint critiquing content that has been updated in the peer-reviewed draft. If the peer-reviewed draft is chosen as the canonical representation of the paper cluster, then the citation context would not accurately capture the rationale of that reference. While worth noting, we believe such cases are rare and do not affect the vast majority of citation contexts.

D S2ORC evaluation criteria

Paper cluster quality For each paper cluster, we compare the selected canonical *Title* and *Authors* fields with the title and authors of the selected canonical PDF. The *Title* field is labeled correct if it exactly matches the title seen on the PDF, with some allowance for different capitalization and minor differences in special character representation (e.g. "γ" versus "gamma") and ignoring whitespace. The *Authors* field is labeled correct if all authors on the PDF are presented in the correct order, with some allowance for variation in the surface form. This is to avoid penalizing publisher metadata for providing a first initial (instead of the first name) or omitting middle names or titles (e.g. "Dr.", "PhD").

Paper-Bibliography linking For each paper-bibliography pair, we compare the selected canonical *Title* and *Authors* fields in the structured bibliography entry to the selected canonical *Title* and *Authors* fields of the linked paper cluster. The *Title* fields are labeled as a match under the same criteria described above for matching paper cluster *Title* fields and PDF titles. The *Authors* fields are labeled as a match if there is substantial overlap in the names of the authors. For example, if authors A, B and C are in the bibliography entry and the linked paper cluster has authors A and B, then this is still considered a match. We note that in our evaluation, differences in the two sets of author names primarily stems from incorrectly written bibliography entries or mistakes in publisher-provided metadata.

E Training corpus sizes for other language models

Language model	Training data
ELMo (Peters et al., 2018a)	1BW (800M) Wikipedia (1.9B) WMT 2008-2012 (3.6B)
BERT (Devlin et al., 2019)	BooksCorpus (800M) Wikipedia (2.5B)
RoBERTa (Liu et al., 2019b)	BooksCorpus (800M) CC-News (~3.8B) OpenWebText (~1.9B) Stories (~1.6B)
GPT2 (Radford et al., 2019)	Web Text Corpus (~2.8B)

Table 9: Reported and estimated (several papers report corpus size in terms of bytes) token counts of training data used to train language models.

We estimate that all of S2ORC consists of approximately 25B tokens of full body text and 15B tokens of abstract text. As demonstrated for S2ORC-SciBERT pretraining, aggressively-cleaned body text from the PDF-parsed subset of S2ORC still yields approximately 16.5B tokens. The size of S2ORC makes it more than sufficient for pretraining large language models such as ELMo, BERT, RoBERTa, GPT2, and others, whose reported training data sizes are given in Table 9 for comparison.

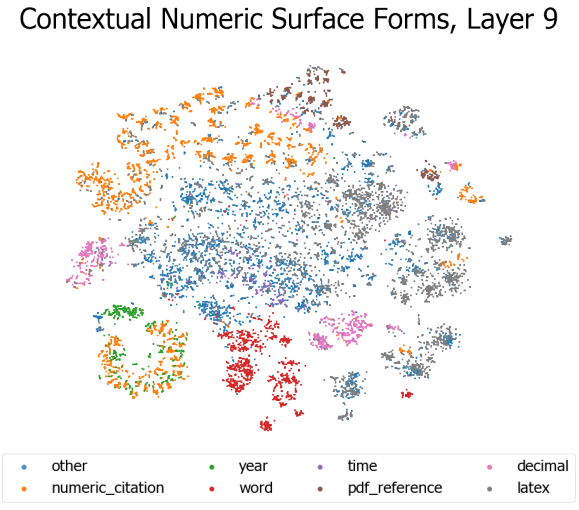

Figure 4: Visualization of contextual representations from layer 9 of S2ORC-SciBERT on numeric surface forms in a subsample of body text from S2ORC. Labels are heuristics based on token-level patterns.

F Numeric representations in S2ORC-SciBERT

Academic papers contain substantially more diverse uses of numeric surface forms than typical web text, such as experimental results, equations, citation references and section/figure markers. To demonstrate this, we cluster contextual word representations involving numbers, heuristically labeling them into one of 8 categories based on surface patterns. Examining the progression of the contextual representations through the layers of BERT reveals an initial focus on sentence position (expected, due to explicit position embeddings) and magnitude, with later layers integrating substantial contextual information, such as the presence of inline LaTeX identifiers, citation indicators and PDF references. Following Peters et al. (2018b); Liu et al. (2019a), we observe that the final 2-3 BERT layers provide embeddings that excel at predictive language modeling; as such, Figure 4 uses embeddings from layer 9 of S2ORC-SciBERT.