MOTIVATION.

[question1] For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
[answer1] The dataset was created as an evaluation set for measuring information retrieval methods for a faceted Query by Example task in scientific papers. The nature of similarity judgments are meant to facilitate development of models which capture relational similarities between aspects (background, method, result) of a scientific papers, these kinds of similarities have been important for creative activities like scientific research. Our paper and `ann_guidelines.pdf` elaborate on similarity guidelines in more detail.
[question2] Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?
[answer2] The dataset was created by researchers at the [Center for Intelligent Information Retrieval](https://ciir.cs.umass.edu/) and the [Information Extraction and Synthesis Laboratory](https://www.iesl.cs.umass.edu/) at the University of Massachusetts Amherst on behalf of those same entities.
[question3] Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.
[answer3]
[question4] Any other comments?
[answer4] No.

COMPOSITION.

[question5] What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
[answer5] The dataset may be viewed from a handful different perspectives: The dataset may be seen as consisting of 50 query abstracts. These are abstracts from scientific papers appearing in the NLP publication venues. Each of the 50 queries have either 250 or 100 other candidate abstracts labeled for similarity with respect to the query along one of 3 "facets". Therefore each of the query abstracts is also paired with one of 3 query facets.
[question6] How many instances are there in total (of each type, if appropriate)?
[answer6] In all, the dataset consists of 6244 query-candidate pairs. All abstracts are also paired with the corresponding paper title and have a potential incomplete set of bibliographic (authors, publication venue and year, doi etc) information.
[question7] Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).

[answer7] The dataset is a sample from a larger set. The query abstracts are drawn from the ACL Anthology. The candidate abstracts are drawn from computer science papers from arXiv which were included in the [Semantic Scholar Open Research Corpus](https://github.com/allenai/s2orc) (S2ORC).

[question8] What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.

[answer8]

[question9] Is there a label or target associated with each instance? If so, please provide a description.

[answer9] Each query-candidate pair is rated on a scale of 0-3 indicating the relevance of the candidate abstract to the query abstract along the query facet. The definitions of relevance are detailed in the `ann_guidelines.pdf`.

[question10] Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.

[answer10] Bibliographic information (`metadata` fields in the `abstracts-csfcube-preds.jsonl` file) for the abstracts was drawn from that present in the S2ORC corpus. This information may be incomplete.

[question11] Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

[answer11] Individual papers (abstracts of which are in the dataset) are part of the citation network, this information is missing from this dataset but it could be obtained from the S2ORC corpus.

[question12] Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

[answer12] The dataset only represents an evaluation set for the tasks it was developed for. Therefore it only consists of development and test splits. We recommend reporting results for each of the facets in the dataset (background, method, result) separately along side a facet combined 'all' split. For each split per facet we follow a 2-fold cross validation approach where half the queries are considered dev and a the other half test, this is done 2 times. Results per query are averaged across the development splits in the 2 folds to obtain the development metric and the scores across the test splits in the
two folds give the test score metric. The splits are illustrated below for a specific facet where q_1 to q_16 represents queries in a list of 16 queries. We recommend using the development split of a single fold (1 or 2) for model development to avoid model development on the test metrics. If
training a single model for all facets for the task we recommend using a single dev fold for the 'all' set of the data which contains the three facets combined into one. For computing statistical significance of results we recommend using metrics on all the queries, per facet and in the 'all' set.

[question13] Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

[answer13] The dataset was built from abstracts, titles, metadata, and the citation network included as part of the S2ORC corpus. Several elements of this corpus were constructed using automatic tools to obtain paper metadata, abstracts, citation span information and so on. This introduces an element of noise in our dataset, for example some candidate abstracts can be noisy (query abstracts were filtered for noise manually). Further the query and candidate abstracts sentences have a label indicating the facet for the sentence, this label was automatically predicted (using this model: [link](https://github.com/allenai/sequential_sentence_classification)) and corrected for the query abstracts but not for the candidate abstracts. Incorrect predictions persist in the dataset.

[question14] Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

[answer14] The dataset is self contained.

[question15] Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

[answer15] No.

[question16] Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

[answer16] No.

If the dataset does not relate to people, you may skip the remaining questions in this section.

[SkipA] NO

[question17] Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

[answer17] No.

[question18] Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.

[answer18] The authors of individual papers included in the dataset are present as part of metadata. If absent, web searches can easily reveal authors.

[question19] Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual

orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
[answer19] No.
[question20] Any other comments?
[answer20] No.

COLLECTION PROCESS.

[question21] How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
[answer21] The dataset is used as is from the S2ORC corpus. Our process for selecting query was a mix of manual curation and automatic selection. `ann_guidelines.pdf` details this process.
[question22] What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?
[answer22]
[question23] If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?
[answer23] While `ann_guidelines.pdf` details the process of collecting the queries and candidate abstracts. The 800,000 computer science papers in this dataset were obtained from the S2ORC corpus as follows: 1. Papers tagged with "Computer Science" in the `mag_fos` field and with a non-null `has_arxiv_id` field in the S2ORC metadata tsv files selected to ensure computer science papers with likely full body text available. This resulted in about 140k papers. 2. For these papers all the outgoing citations which are part of the S2ORC corpus were obtained. This resulted in about 1.2 million papers. 3. For these papers any of the following filters based on sentences (nltk.tokenize.sent_tokenize) and tokens (white space split the sentence) evaluating to
true excludes the paper: abstract has fewer then 3 sentences, abstract has greater then 20 sentences, any sentence is greater than 80 tokens, or all sentences have fewer than 4 tokens. This procedure results in about 800,000 abstracts based on which the remainder of the corpus is built.
[question24] Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?

[answer24] One graduate student, a post doc and 2 hired annotators (both graduate students) were involved in creation of the dataset. Both hired annotators were paid $22.5/hour for 25 hours of work for a period of 3 weeks.

[question25] Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

[answer25] Initial rounds of exploratory annotation were carried out over October 2020-December 2020. The dataset released here was annotated from January 2021-February 2021.

[question26] Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

[answer26] The release of demographic information of annotators was reviewed by the University of Massachusetts Amherst Human Research Protection Office (HRPO) and determined as not meeting the definition of human subjects research (and hence exempt from IRB review). No other elements of the project were reviewed by research ethics compliance bodies.

If the dataset does not relate to people, you may skip the remaining questions in this section.

[SkipB] NO

[question27] Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?

[answer27] Scientific papers were gathered from the S2ORC corpus. Relevance judgments were made directly by individual annotators

[question28] Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

[answer28] No.

[question29] Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

[answer29] N/A

[question30] If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

[answer30] No.

[question31] Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If

so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
[answer31] No.
[question32] Any other comments?
[answer32] No.

PREPROCESSING/CLEANING/LABELING.

[question33] Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.
[answer33] Aside from the data filtering for excluding noisy data described above no other pre-processing was applied on the data.
[question34] Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
[answer34] N/A
[question35] Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.
[answer35] This will be released in future.
[question36] Any other comments?
[answer36] No.

USES.

[question37] Has the dataset been used for any tasks already? If so, please provide a description.
[answer37] No.
[question38] Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
[answer38] No.
[question39] What (other) tasks could the dataset be used for?
[answer39] The dataset is intended for the two formulations of the faceted Query by Example task as described in the paper accompanying the dataset. Additionally it is conceivable this could be used for evaluating methods which diversify retrieved papers along different facets (using the 16 query papers which have been rated for similarity along 2 different facets each). The dataset can also be used to evaluate a range of other methods which present methods of measuring general scientific paper similarity.
[question40] Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future

uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?
[answer40] None that we are aware of.
[question41] Are there tasks for which the dataset should not be used? If so, please provide a description.
[answer41] None that we can think of.
[question42] Any other comments?
[answer42] No

DISTRIBUTION.

[question43] Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
[answer43] Yes
[question44] How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
[answer44] GitHub. Please use the appropriate [release](https://github.com/iesl/CSFCube/releases) to download salient releases of the dataset.
[question45] When will the dataset be distributed?
[answer45] The dataset has been publicly available since March 3rd 2021.
[question46] Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
[answer46] The dataset is released under the [Creative Commons Attribution-NonCommercial 4.0 International](https://creativecommons.org/licenses/by-nc/4.0/legalcode) license.
[question47] Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
[answer47] None
[question48] Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

[answer48] None
[question49] Any other comments?
[answer49] No

MAINTENANCE.

[question50] Who will be supporting/hosting/maintaining the dataset?
[answer50] Sheshera Mysore (smysore@cs.umass.edu)
[question51] How can the owner/curator/manager of the dataset be contacted (for example, email address)?
[answer51] Yes.
[question52] Is there an erratum? If so, please provide a link or other access point.
[answer52] None yet.
[question53] Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?
[answer53] If sufficiently large errors are discovered the dataset will be corrected and updated version of it will be released. We will use the "Releases" feature on Github to denote all salient releases.
[question54] If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
[answer54]
[question55] Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.
[answer55]
[question56] If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.
[answer56]
[question57] Any other comments?
[answer57] No.