CSFCube – A Test Collection of Computer Science Research Articles for Faceted Query by Example

Sheshera Mysore[1] **Tim O'Gorman**[2]* **Andrew McCallum**[1] **Hamed Zamani**[1]

{smysore, mccallum, zamani}@cs.umass.edu

[1]University of Massachusetts, Amherst, MA, USA
[2]Thorn, CA, USA

Abstract

Query by Example is a well-known information retrieval task in which a document is chosen by the user as the search query and the goal is to retrieve relevant documents from a large collection. However, a document often covers multiple aspects of a topic. To address this scenario we introduce the task of *faceted Query by Example* in which users can also specify a finer grained aspect in addition to the input query document. We focus on the application of this task in scientific literature search. We envision models which are able to retrieve scientific papers analogous to a query scientific paper along specifically chosen rhetorical structure elements as one solution to this problem. In this work, the rhetorical structure elements, which we refer to as *facets*, indicate objectives, methods, or results of a scientific paper. We introduce and describe an expert annotated test collection to evaluate models trained to perform this task. Our test collection consists of a diverse set of 50 query documents in English, drawn from computational linguistics and machine learning venues. We carefully follow the annotation guideline used by TREC for depth-k pooling (k = 100 or 250) and the resulting data collection consists of graded relevance scores with high annotation agreement. State of the art models evaluated on our dataset show a significant gap to be closed in further work. Our dataset may be accessed here: https://github.com/iesl/CSFCube

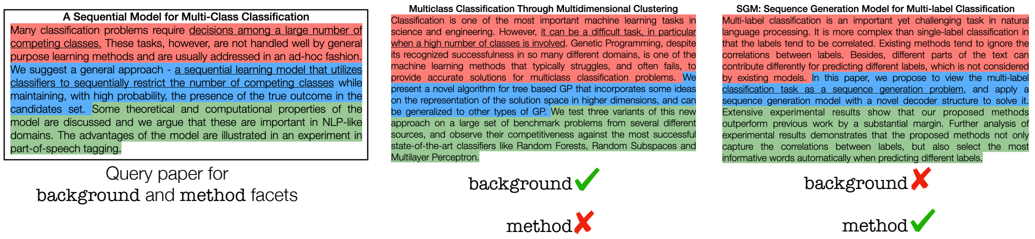
1 Introduction

The dominant paradigm of information retrieval is to treat queries and documents as different kinds of objects, e.g., in keyword search. This paradigm, however, does not lend itself to exploratory search tasks. On the other hand, paradigms of search such Query by Example (QBE) which treat queries and documents as similar kinds of objects have been considered more suited to exploratory search tasks [41, 17, 45]. QBE has also been used more recently in information extraction for NLP [61, 58], and seems poised to leverage recent advances in representation learning and NLP for search tasks where keyword search proves insufficient [6, 70].

In QBE search tasks, each query is a document that often covers multiple aspects of a topic leading to a document-only query under-specifying how retrievals should me made. Here, we introduce the task of faceted QBE, where users can specify an information need by providing an input document and a facet example, with the goal to retrieve documents that are similar to the input document from the perspective of the given facet. This paper focuses on the case of faceted QBE applied to scientific articles. Figure 1 illustrates how multi-aspect similarities may arise in scientific articles, where candidate documents could be similar to the query along the general problem being addressed or the method used in a paper.

*Work done while at the University of Massachusetts, Amherst.

35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks.



A Sequential Model for Multi-Class Classification

Many classification problems require decisions among a large number of competing classes. These tasks, however, are not handled well by general purpose learning methods and are usually addressed in an ad-hoc fashion. We suggest a general approach - a sequential learning model that utilizes classifiers to sequentially restrict the number of competing classes while maintaining, with high probability, the presence of the true outcome in the candidates set. Some theoretical and computational properties of the model are discussed and we argue that these are important in NLP-like domains. The advantages of the model are illustrated in an experiment in part-of-speech tagging.

Multiclass Classification Through Multidimensional Clustering

Classification is one of the most important machine learning tasks in science and engineering. However, it can be a difficult task, in particular when a high number of classes is involved. Genetic Programming, despite its recognized successfulness in so many different domains, is one of the machine learning methods that typically struggles, and often fails, to provide accurate solutions for multiclass classification problems. We present a novel algorithm for tree based GP that incorporates some ideas on the representation of the solution space in higher dimensions, and can be generalized to other types of GP. We test three variants of this new approach on a large set of benchmark problems from several different sources, and observe their competitiveness against the most successful state-of-the-art classifiers like Random Forests, Random Subspaces and Multilayer Perceptron.

SGM: Sequence Generation Model for Multi-label Classification

Multi-label classification is an important yet challenging task in natural language processing. It is more complex than single-label classification in that the labels tend to be correlated. Existing methods tend to ignore the correlations between labels. Besides, different parts of the text can contribute differently for predicting different labels, which is not considered by existing models. In this paper, we propose to view the multi-label classification task as a sequence generation problem, and apply a sequence generation model with a novel decoder structure to solve it. Extensive experimental results show that our proposed methods outperform previous work by a substantial margin. Further analysis of experimental results demonstrates that the proposed methods not only capture the correlations between labels, but also select the most informative words automatically when predicting different labels.

Figure 1: Examples of candidate papers similar along different facets for the same query paper, underlined text indicates similar aspects: A paper discussing a classification problem with large number of classes is similar along `background` with a paper discussing the same problem, but with a different paper discussing a sequential model for its `method` facet.

The promise of literature search to navigate vast collections of research documents offers exciting prospects and the potential to accelerate the process of scientific discovery [24]. This optimism has been expressed in a range of work in biomedicine [67], materials science [27], geography [42], biomimetic design [40], and machine learning [66]. The ever-growing research literature has also lead to search and recommendation tools being important in the workflows of individual researchers across a range of disciplines [43, 55]. However, few literature search and recommendation tools address the problem that documents often contain multiple fine grained aspects one might want to search with [53]. Further, prior work has also demonstrated that researchers often desire finer-grained control in literature search and exploration tools [18, 51, 28]. Context-dependent faceted search tools have the potential to fill this gap. Finally, while we focus on literature search, faceted QBE also has the potential to serve in other applications such as e-commerce search and product design [29].

The problem at the heart of faceted QBE is that of fine-grained document similarity. This is of broader relevance to numerous other research problems. This work therefore promises to be of value in research problems such as that of multi-aspect reviewer assignments for papers [35], tracing the adoptions of ideas from the research literature [11], or that of making causal inferences with matched scientific texts [57, Sec: Applications and Simulations]. Appendix E elaborates on applications.

Despite this wide applicability of the problem, and a range of proposed approaches, evaluation of these document retrieval/similarity methods remains a problem – with prior work relying on weak sources of gold evaluation data or on expensive human evaluations. A broad set of approaches evaluate systems against citations or combine citation information with other incidental information such as section headers as a means to determine citation intents acting as proxies for facets [54]. While these sources may represent reasonable noisy signals [33, Sec 4] for model training, the noise in such approaches limits their value as a method of system evaluation (Appenxix D presents an analysis of citations as evaluation data). Other work has relied on extrinsic evaluations in a downstream ideation task [13] or based on feedback of users interacting with a recommendation system [12, 1]. However, such approaches require expensive evaluation protocols unsuitable for model development and system comparisons. Therefore, we believe our dataset will fill an important gap by presenting a pragmatic alternative to these extremes.

Our publicly accessible test collection consists of 50 diverse research paper abstracts in English paired with a facet as queries (§3). The query facets are chosen from one among three facets from $\mathcal{F} = \{\texttt{background/objective}, \texttt{method}, \texttt{result}\}$, representing the three dominant aspects of methodological scientific research. Query abstracts are selected from the domains of machine learning and natural language processing. Candidate pools of size 250 (8 queries) or 100 (42 queries) are constructed from about $800,000$ computer science papers in the S2ORC Corpus [46] using a diverse set of retrieval methods. Four graduate-level computer science annotators with research experience rated each candidate with respect to the query abstract and facet, with high agreement. Finally, we also present a range of baseline experiments and analysis (§5) indicating the phenomena that state of the art models fail to capture, presenting significant room for improvement.

2 Task Formulations

We frame our task as one of retrieving scientific papers given a query paper and additional information indicating the query facet. In this work we operate at the level of abstracts rather than the body text of papers under the understanding that salient information about a paper is contained in the abstract, especially for the domains considered in this paper [30]. Further, important applications in large deployed systems for paper recommendation and reviewer assignment operate at the level of paper abstracts indicating them to be an already useful choice [60, 52]. Our decision is also motivated by the difficulty of annotating large corpora at the body-text level, and mirrors a common choice in numerous prior work [67, 10]. Note however, that our dataset is structured to leverage future full-text approaches (See §3). Finally, §4 presents a small-scale analysis of full-text vs abstract annotation consistency. In what follows, we also assume access to the title of the papers, although we drop it from the below discussion for brevity.

We denote a query document with Q, a candidate document $C \in \mathcal{C}$. Every Q or C, consist of N sentences $\langle S_1, S_2, \ldots S_N \rangle$. We denote *facets*, f from an inventory of labels $\mathcal{F}$, indicating the rhetorical elements of the document. We denote a ranked list of the candidates for Q and query facet f_q with $\langle r, C \rangle \in R_{Qf}$ where elements denotes document C at rank position r. Now, we formulate two tasks that our test collection will effectively evaluate:

Definition 1. *Retrieval based on pre-defined facets: Given query and candidate documents – Q and $\mathcal{C}$, with sentences in both annotated with facet labels: $\langle (f_1, S_1), (f_2, S_2), \ldots (f_N, S_N) \rangle$ and a query facet f_q a system must output the ranking R_{Qf}.*

Definition 2. *Retrieval by sentences: Given query and candidate documents – Q and $\mathcal{C}$, and a subset of sentences $\mathcal{S}_Q \subseteq Q$ based on which to retrieve documents a system must output the ranking R_Q.*

Def 1 corresponds most closely to faceted search as described by Kong [37] and closest to work by Chan et al. [13]. While Def 2 represents a more general formulation not relying on pre-specified facet labels. Here, the sentences $\mathcal{S}_Q$ can be viewed as exemplar facet sentences based on which results must be retrieved even while lacking any explicit facet specification.

One may view both Definition 1 and 2 as instances of the QBE paradigm of retrieval [45]. One, at the level of documents, using the document Q as a query as in El-Arini and Guestrin [19] and Zhu and Wu [70], and a second at the level of sentences denoting a facet of the paper. Broadly, we believe QBE to be well suited to the problem of faceted literature search given the difficulty of being able to specify in keyword searches precisely the search intent and given that the meaning of sentences denoting a facet are often dependent on the broader context of the abstract. Further, we expect Definition 2 to have specific other advantages: users often tend to have different understanding of facets than those defined by designers of the ranking system [64] – in our case we expect that different sub-areas/areas of the literature will exemplify different kinds of facets making it hard to pre-specify facets in a system. Further, users often wish to explore the literature at different levels of granularity than that possible with pre-defined facets [28, page 14], we expect QBE will allow users greater control to select parts of an abstract expressing a facet at different levels of granularity based on how they would like papers retrieved. Importantly however, note that while our dataset selects a specific set of facets for ease of annotation it facilitates evaluation and consequent model development of both task setups.

3 Dataset Description

In the construction of this test collection we relied on the Semantic Scholar Open Research Corpus (S2ORC) [46] which provides a corpus of 81.1M English language research papers alongside a range of automatically generated metadata including citation network information. We choose to work with about $800,000$ computer science papers in S2ORC sourced from arXiv.[2] These papers were selected to ensure that the full-body text of the papers was available, in addition to the abstract and title, to facilitate potential future research.[3] Our queries were selected from domains of machine learning and NLP so that annotators would be familiar with the domain in question.

[2] `https://arxiv.org/`

[3] `datasheet.md` in the dataset release documents detailed filtering steps used to obtain the 800,000 documents. Our release also includes these 800,000 documents.

Facets: In this work a facet for a research paper corresponds to the dominant steps involved in carrying out scientific research – the identification of a research problem/question (`background/objective`), formation and testing of the hypothesis (`method`), and formation of conclusions (`results`). These facets are broadly defined as:

`background/objective`: Most often sets up the motivation for the work, states how it relates to prior work and states the problem or research question being asked. Henceforth, we refer to these as `background` facets.

`method`: Describes the method being proposed or used in the paper. The method could be described at a very high level or it might be specified at a very fine-grained level depending on the type of paper. Note that our definition of methods is broad and will include methods of analysis of a phenomena, a model, data, or procedural descriptions of the experiments carried out. The specific interpretation of method also depends on the type of paper (§3).

`result`: This may be a detailed statement of the findings of analysis, a statement of results or a concluding set of statements based on the type of paper.

Our corpus is labelled with facets predicted using the model of Cohan et al. [15] into the set of labels: {`background`, `objective`, `method`, `result`, `other`}. Incorrect facet labels for the query abstracts are manually corrected. Prior to relevance annotation, `objective` and `background` are merged into one facet called `background` as they were too similar to be distinguished for the purpose of document similarity. The `other` facet is not considered for annotation. These sentence facets were then provided as additional guidance during annotation, with query facet sentences being bolded to encourage attention to those parts of the document.

Query Abstract Selection: We annotated a total of 50 query abstract-facet pairs from the ACL Anthology.[4] Of the 50, we annotated 16 abstracts with two different facets each (total of 32 query abstract-facet pairs), in order to allow closer analysis of the differences in retrieval performance for the same query abstract while varying the query facet (Figure 3). The remaining 18 abstracts were annotated for a single query facet each. In total, our dataset contains 16 `background` queries, 17 `method` queries, and 17 `result` queries; further statistics are provided in Table 1. Queries were selected to ensure coverage over a range of years (1995-2019) and to ensure a somewhat even distribution across query paper types, as coarsely divided into "resource/evaluation papers", "data-driven approach papers" or "theoretical papers". This was ensured through randomly sampling a set of 100 articles over the time period and were manually filtered to ensure that each query corresponded to specific and non-trivial representations of multiple facets. We include the query data distributions and the procedure for query selection in the annotator guidelines in our dataset release.[5]

Candidate Pooling: Candidates per query are drawn from a corpus of about 800,000 computer science papers in the S2ORC corpus using the following pool of methods: TF-IDF, averaged `word2vec` embeddings, and TF-IDF weighted `word2vec` based similarities run on titles, and abstracts giving us a set total of 6 methods. Further, a state-of-the-art BERT model, SPECTER [16], trained for scientific paper representation using citation network signals was also part of the set of methods used to generate our pool. Finally, papers cited in the query paper are also added to the pool, given their likely relevance due to authors self selections. This set of methods represents a diverse range of similarities with each of the methods retrieving largely different candidate abstracts: The top-25 papers across retrieval methods contained between 1-4 papers in common. For a set of 8 abstract-facet queries, we annotate pools of size 250. These formed an initial exploratory set of annotations, and the remaining 42 queries were annotated with pools roughly of size 100. For queries with pools of size 250 we draw the top 33 papers from each retrieval method, similarly for queries with pools of 100 we draw the top 13 papers. The order in which we draw from the group of methods is randomized for every query and in the case of a candidate already present in our candidate pool we draw from further down the ranked list of a method. Finally, while our task is framed as facet dependent, we use non-faceted methods in the construction of our pool due to the lack of well-established faceted retrieval models. This choice also allowed us to ensure an identical pool of candidates for being annotated with respect to different facets – providing for a richer evaluation setup. Statistics of the dataset are provided in Table 1.

[4] `https://www.aclweb.org/anthology/`

[5] Annotator guideline and query metadata files in our dataset release: `ann_guidelines.pdf` and `queries-release.csv`

Table 1: Statistics for the test collection.

Statistic		All
Query abstract-facet pairs	-	50
Unique query abstracts	-	34
Mean candidate pool size	-	124.9
Query-candidate pairs	-	6244
Candidates rated +1 per query	min	12
	max	87
	avg	36.9
Candidates rated +2/+3 per query	min	1
	max	35
	avg	9.8

Table 2: Spearman's ρ, Krippendorff's α, Cohen's κ, and % agreement measures for relevances before and after the adjudication stage of annotation. Given the ranking nature of the task, Spearman's ρ presents the most apt measure of agreement.

	pre-adjudication			
facet	ρ	α	κ	%
background	0.45	0.43	0.28	57.07
method	0.31	0.26	0.20	69.60
result	0.42	0.35	0.26	67.46
	post-adjudication			
background	0.73	0.72	0.62	77.68
method	0.63	0.61	0.54	84.47
result	0.70	0.67	0.59	83.53

4 Dataset Annotation

Relevance Ratings: We choose to rate candidate documents on a graded scale from 0-3 with our definitions for the scales depending on the facet. Broadly, we train annotators to rate structural/relational similarities between the candidate and query higher (3-2 ratings) than attribute/feature based similarities (1-0 ratings). This draws on motivations from a range of literature highlighting the importance of structural similarities between ideas to creative activities like scientific research – a focus of this work [47, 23, 13, 44]. We illustrate the definitions with the `method` facet here:

`Near Identical/+3`: +3 implies methods described share a similar over-arching mechanistic similarity, further the methods must also be similar in terms of the details of the objects being manipulated.

`Similar/+2`: +2 implies that the methods are mechanistically similar and the details are only comparable between the query and candidate.

`Related/+1`: +1 is meant to encompass a wide range in being similar and can be hard to list at length. Common cases include: 1. Details of the two methods are similar but there is only high level mechanistic similarity. 2. Small or not-so-important mechanistic parts of the methods in two papers are similar. 3. Where query and candidate abstracts may vary in the level of granularity in which they describe a method and a high level similarity is the only one you can establish by reading the abstract.

Our relevance grades also include an `Unrelated/0` grade for documents deemed unrelated. We encourage readers to examine `ann_guidelines.pdf` in our dataset release which details these further alongside examples for every case.

Annotation Procedure: Our annotation was carried out by four graduate-level computer science annotators (the lead authors, SM, TO, and 2 hired annotators) with experience reading research papers in the selected domains.[6] The annotation guidelines were developed over 4 iterations of repeated annotation and refinement of the guidelines. The hired annotators were trained prior to annotation and demonstrated Spearman Correlation based agreements of $0.5 - 0.7$ with an adjudicated training set of examples. Annotators were paid an hourly wage of USD 22.5 for a period of 3 weeks. All query-candidate pairs were annotated by two independent annotators and a third adjudicator resolved the cases of difference between the first two annotators. All annotations were carried out in Label Studio[7] and annotators were only shown paper titles and abstracts at all stages of annotation, hiding all other metadata including authors, publication venues, and years.

Relevance Analysis: Given our two stage annotation process of gathering double annotation and an adjudication stage, we report agreement metrics for both the stages. Table 2 presents these agreements. Our pre-adjudication metrics are those between the two annotators involved in the annotation. For the post-adjudication metrics, we report the mean metrics between the adjudicated ratings and each of the two initial annotators ratings. We report Spearman rank-correlation coefficients, ρ, between

[6] Further details about annotators are included in the `datasheet.md` in the dataset release.

[7] `https://labelstud.io/`

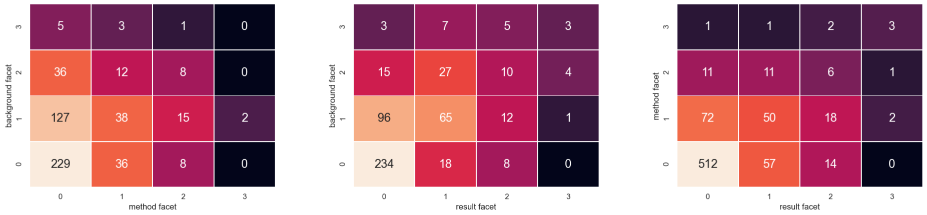


Figure 2: Number of candidates labeled with a particular facet relevance (scale of 0-3) for the documents labeled with two facets. Note that while some between facets correlation exists (along the diagonals), many candidates are relevant only to one facet. Also note that this differs by facet pair.

annotators (pre-adjudication) and between annotators and adjudicated rankings (post-adjudication) produced by ratings. We believe ρ to be the most apt measure of agreement given our ranking task, where we are most interested in establishing a relative similarity between papers. Our reporting also follows work in rating sentence similarities which reports correlations between annotators as a measure of agreement [3]. Additionally, we report Krippendorff's α with an ordinal distance function to measure agreement in absolute terms while taking into account an ordinal relevance scale, Cohens κ to measure agreement while not taking into account the ordinal nature of relevance levels i.e. treating ratings as categorical labels, and simple percent agreements as an illustration of the fraction of data which needed adjudication. All of the metrics in Table 2 represent median values across the per-query metrics. It was also permissible for the adjudicator to entirely over-rule both annotators ratings for a candidate. Across all facets, we saw this happen very rarely, 2-3% on average per query.

Based on the observed values of ρ, and instances of over-ruling in adjudication we believe that annotators are able to consistently establish a relative similarity between papers and indicate strong agreement with an adjudicated set of ratings. Based on qualitative observations in the annotation process we noted the primary case of disagreement between annotators. Disagreements occurred most where a single facet was representative of multiple different finer-grained aspects. In these cases annotators initially focus on one of the aspects in making their annotations, when made aware of other aspects during adjudication, annotators readily accepted a different judgement. Appendix G presents an example. We believe this speaks to the effectiveness of an adjudication step. Finally, Figure 2, indicates the set of relevances for query abstracts annotated with two facets – we see that while some candidates are correlated in relevance others are only relevant to the query in one facet.

Full-text vs abstract annotations: To examine the effect of annotating the abstract of papers instead of full-text of papers we also conduct a small scale study to examine the extent of differences between the relevances produced by them. This is done by a single expert annotator annotating relevance based on abstract and full-text separately for 9 query abstracts (3 from each facet) and 5 candidate abstracts each (45 pairs). We refer readers to Appendix B for the details of the annotation setup. Agreement was measured between the relevances produced based on the abstract text and the full-text: Spearmans $\rho = 0.78$, Krippendorff's $\alpha = 0.77$, Cohens $\kappa = 0.63$, and $\%$ agreement of 73%. Full-text annotation was performed at the rate of about 5 minutes/pair, on the other hand abstract ratings took 15 seconds/pair. While our metrics indicates a less than perfect match between the abstract ratings vs full-text these metrics indicate strong agreement between the two. We believe this presents a reasonable trade-off between expense and completeness of annotation. Further, prior work on a similar scientific similarity annotation task noted low-agreements between expert annotators in annotation of full-texts [68, Sec 7], indicating that use of full-text does not necessarily translate into higher quality annotations.

Comparison to Other Datasets: To the best of our knowledge, Neves et al. [51] and Chan et al. [13] present the only other datasets of similar structure to the one we present:

Neves 2019: Neves et al. [51] annotate a set of 8 query papers with 70-90 candidates per query from biomedicine for similarity of the "papers research goal".

SOLVENT: Chan et al. [13, Sec 4.1] annotate a small scale dataset of 50 social computing papers for similarity along "purpose" or "mechanism" facets. Here, each of the 50 papers is annotated for relevance with respect to the other 49 papers in the set of 50 papers.

Both these datasets intend at presenting small scale evaluations in the context of other work. Our work, in contrast, presents a substantially larger resource created over multiple annotation passes, annotates similarity across multiple facets, and presents a reusable dataset. Several other datasets bear partial resemblances to the work we present here, we discuss these in Section 7.

Annotation scalability: We believe that a dataset like ours necessarily calls for expert annotation. Understanding the annotation guidelines alone required experience reading research papers. We believe these traits make it harder to scale the presented datasets using an un-trained crowd-sourcing based method. However, we believe certain aspects of our work can inform future work to scale datasets like ours. A dominant approach in building datasets involves annotation of instances by multiple annotators followed by a majority vote [9] or an average [3] to determine the "gold" label. Given the expense of creating repeated annotations when working with expert annotators we instead choose to use a adjudication step to create "gold" labels. We believe this helped us scale our approach considerably. If this step leveraged more experienced annotators to perform adjudication while lower experience annotators make initial judgements we believe it would be scalable beyond the approach presented here. This also follows prior work from Nallapati et al. [49], which leverages a hierarchy of annotator skills to scale a complex annotation task. Finally, we also note that large IR test collections call for competing systems to generate a pool of documents to judge [67]. We believe our dataset presents a robust evaluation set which may be leveraged for development of such methods.

5 Experimental Results

Here, we establish baseline performances from a handful of standard and state of the art methods.

Baseline methods: The methods we choose to evaluate capture a range of granularities and nature of methods: term based methods, pre-trained model based sentence representations, and whole abstract representation models. Appendix C describes each method in detail.

Term-level baselines: `fabs_tfidf`, `fabs_bm25`, `fabs_cbow200`, and `fabs_tfidfcbow200` represent term-level baselines. These represent the document as sparse TF-IDF vectors, averaged bag of word representations, and weighted averaged bag of word representations respectively. Each of them represent the query document as the representation for the sentences corresponding to the query facet sentences in the query abstract. Candidates are represented by their whole abstract representations.

Sentence-level baselines: Here encode all query facet sentences and all candidate abstract sentences individually with a sentence encoder, and then use the maximum pairwise sentence cosine similarity between the query and candidate sentences to rank candidates. `SentBERT-PP`, `SentBERT-NLI`, `UnSimCSE`, `SuSimCSE` (unsupervised and supervised `SimCSE`) represent state of the art sentence encoder baseline models [56, 22]. Each of the models here represent models trained in a different manner or on different training sets.

Abstract-level baselines: SPECTER and SPECTER-ID represent whole abstract level representations [16]. Both of these approaches represents a multi-layer transformer based SciBERT model fine-tuned on citation network data. SPECTER operates on titles and the whole abstract of the papers. Both queries and candidates are represented by their SPECTER embeddings. Note that SPECTER was trained on a corpus of randomly selected scientific documents. We also re-implement and train a version of SPECTER on about 660k computer science paper triples with identical hyper-parameters to SPECTER, we call this in-domain model SPECTER-ID.

In re-ranking we use the L2 distance between the query and candidate vectors unless noted otherwise. Note here that while the term-level baselines are more similar to the the task formulated in Definition 1, the sentence-level baselines solve the task in Definition 2. Further note that we make sure to include baseline methods from the above method types such that were not used for the constructions of pools. This is intended to evaluate the performance of methods which were not used for pool construction (§3), thereby investigating the ability of the dataset to be re-used for evaluating future methods. These are represented by `fabs_bm25`, `SentBERT` and `SimCSE` based methods, and the SPECTER-ID baseline.

Re-ranking Results: Table 3 denotes performance on the test set for each facet independently and aggregated in the *Aggregated* columns. In reporting results, we report R-Precision, Precision@20, recall@20, and NDCG@k. For NDCG@k, we follow Wang et al. [69], and choose $k = p * |\mathcal{C}|$ where

Table 3: Test set results for the set of baselines methods. Metrics (R-Precision, Precision and Recall at 20, NDCG$_{\%20}$) are computed based on a 2-fold cross-validation, represent averages over per-query metrics, and are reported as percentages. SPECTER-ID performance is reported over three training re-runs with underset standard-deviation, the remaining baselines are reported based on a single set of model parameters released by the respective authors.

	background				method			
	RP	P@20	R@20	NDCG$_{\%20}$	RP	P@20	R@20	NDCG$_{\%20}$
fabs_tfidf	23.35	27.19	45.80	57.97	09.30	09.83	34.75	31.20
fabs_bm25	20.12	27.81	49.85	59.39	09.37	11.63	38.29	34.59
fabs_cbow200	19.61	15.94	27.97	36.56	08.65	08.33	15.69	21.14
fabs_tfidfcbow200	15.92	16.87	27.76	40.51	07.99	06.01	17.71	21.70
SentBERT-PP	21.24	28.75	46.67	60.80	10.00	10.83	36.30	33.40
SentBERT-NLI	19.02	25.00	40.13	54.23	09.11	11.46	02.89	31.10
UnSimCSE-BERT	18.15	23.44	36.05	51.59	08.86	09.65	27.92	31.23
SuSimCSE-BERT	19.22	22.81	46.75	55.22	08.58	09.76	29.01	30.88
SPECTER	**24.81**	**35.31**	**57.45**	66.70	**11.72**	13.58	40.81	37.41
SPECTER-ID	24.55 ±1.3	34.17 ±0.5	53.26 ±0.3	**69.22** ±1.71	10.53 ±0.3	**16.22** ±1.21	**44.59** ±3.6	**42.76** ±0.78

	result				*Aggregated*			
	RP	P@20	R@20	NDCG$_{\%20}$	RP	P@20	R@20	NDCG$_{\%20}$
fabs_tfidf	11.35	16.28	38.57	41.24	14.59	17.64	39.69	43.19
fabs_bm25	11.31	20.00	40.40	45.07	13.50	19.69	42.73	46.06
fabs_cbow200	11.16	10.42	23.44	30.93	13.08	11.47	22.23	29.36
fabs_tfidfcbow200	10.43	10.69	24.39	32.79	11.38	11.09	23.13	31.42
SentBERT-PP	13.60	19.83	41.73	52.35	14.83	19.62	41.41	48.57
SentBERT-NLI	14.23	22.05	46.99	51.30	14.04	19.42	38.67	45.39
UnSimCSE-BERT	12.00	19.58	38.95	45.55	12.92	17.41	34.43	42.59
SuSimCSE-BERT	12.37	18.58	39.76	44.93	13.33	16.95	34.83	43.45
SPECTER	18.62	23.78	52.72	56.67	18.29	23.97	50.14	53.28
SPECTER-ID	**20.09** ±0.92	**27.36** ±0.45	**58.74** ±3.04	**60.40** ±1.31	**18.32** ±0.79	**25.74** ±0.22	**52.12** ±1.54	**57.22** ±0.70

$p \in (0, 1)$. NDCG$_{\%20}$ therefore refers to NDCG computed at 20% of the pool size for a query. This is apt since our queries don't have identical pool sizes. Appendix F presents an extended result table.

Overall results: First, in line with the strong performance of pre-trained language model representations for a range of tasks, we note broadly the stronger performance of SPECTER-ID and SentBERT models compared to term and static embedding based baselines. However, note that the sentence level SentBERT models underperform models which incorporate the whole abstract context. Also note that training on in-domain data allows SPECTER-ID some gains over SPECTER. In examining facet dependent performance, we note the stronger performance of all the methods on the background facet, which is expected due to the stronger correlation between background sentences and the general topic of the paper. Next, we note the consistently poorer performance of all methods on the method facet, providing clear room for improvement. As might be noted from our relevance rating guidelines (§4), we rate "mechanistic" notions of similarity for the method facet. Given this relational nature of similarity, we expect methods relying on whole paragraph or term level representations to perform poorly on this facet. We note results midway between the other facets for result – this is due to some results being easy to be judged similar based on term overlaps while others require deeper a understanding of the query (See Appendix G). Finally, we make special note of the poor performance of recent state of the art models SimCSE, and SPECTER on overall performances, specially so in method and result facets – we believe this offers future work substantial room for improvement.

Since we annotate multiple queries per abstract we also present per-query results for the best performing baseline, SPECTER-ID, on this set of queries in Figure 3. We note here the difference in performance by facet for SPECTER-ID, an un-faceted model. Here, performing well on one facet does not always lead to strong performance on other facets indicating room for improvement with models which incorporate finer-grained conditional similarities into their rankings.

Reusabilty: Since we evaluate methods not used for pool construction (i.e fabs_bm25, SentBERT and SimCSE methods, and SPECTER-ID), we examine their performance. Here, both fabs_bm25 and SPECTER-ID outperform corresponding methods of their method-type used for pool construction (i.e. fabs_tfidf and SPECTER). Further SentBERT-PP, a method representing a different class of method than those used for pool construction also outperforms the kinds of methods used for pool construction, notably fabs_tfidf. We believe this indicates the lack of a serious bias of the dataset toward methods used for pool-construction, allowing re-use for evaluating future methods.

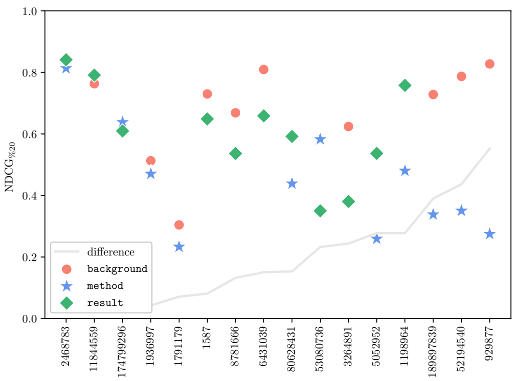


Figure 3: Per-query performance (NDCG$_{\%20}$) for SPECTER-ID on the set of papers which have been annotated with two facets (§3). Performances are means across 3 separate model training re-runs. The query papers are sorted by difference in performance on the facets.

6 Error Analysis

Based on a qualitative examination of per-query ranking performance of `abs_tfidf`, `SentBERT-PP` and SPECTER-ID we outline the factors which lead the baseline models to underperform. We believe the incorporation of modeling to handle these phenomena will lead to improved performance on the task. While we provide a summary of here, Appendix G provides more extensive examples.

Salient Aspects: One source of error is the inability of models to identify the most salient aspects for similarity – representations capturing "what the paper is about?", often expressed only in part of a larger set of facet sentences.

Multiple Aspects: Within a given facet, papers often expressed multiple finer grained aspects, models however often only retrieved based on a single aspect.

Domain specific similarities: A set of errors also arise from the inability of models to determine similarity between technical concepts. For example, an inability to judge "stacking", "ensemble strategy", and "bagging" as similar.

Mechanistic similarities: Nearly all methods perform poorly in the case of determining mechanistic similarity in `method` facets. This often relies on determining similarity across a sequence of actions. This problem also bears resemblance to the challenging setup of retrieval based on movie plots as in Arguello et al. [6].

Context dependence of facets: Faceted similarities as labelled here often also show context dependence on other facets, especially for `result` queries. Given that one major guideline for result similarity in our dataset is "the same finding or conclusion", being able to determine context similarity is important.

Qualitative result statements: `result` queries which summarize qualitative findings as, opposed to reports of performance on a dataset, often perform poorer, often requiring broader context and often lacking in term overlaps which may otherwise easily indicate relevance.

Therefore, the range of challenging phenomena captured in our dataset allow evaluation for novel modeling approaches to document similarity which in-turn translate into progress on a range of important problems. Appendix H indicates potential sources of training data which may be leveraged to train fine-grained document similarity models to overcome some of the above challenges.

7 Related Work

Work presented here ties to that of information retrieval and scientific literature search communities.

Faceted Search: IR has considered the task of faceted search, where facets have often been treated as fixed attributes of metadata [38], in line with a QBE setup our work provides a more semantic interpretation of a facet tied to the rhetorical structure of the document. Other similar work in IR comes from those aiming to diversify search results along the specific aspects/intents of an underspecified keyword query [4]; one might consider the "aspects" in this line of work to our "facets".

Others have also explored dynamic generation of facet like attributes for queries [39]. Partly similar to our task, Upadhyay et al. [65], allow specification of aspects along side ad-hoc queries.

Query-by-Example: A range of literature has also considered the QBE formulation applied to a variety of different kinds of data, from graphs, text, music and to archival image search [45, 2, 34]. Both Sarwar and Allan [59] and Taub Tabib et al. [62] frame retrieval from text corpora using event or syntactic representations as QBE. The closest work in QBE applied to research papers search comes from El-Arini and Guestrin [19], Jacovi et al. [31] which considers multiple document queries for research paper recommendation and Zhu and Wu [70], who additionally propose considering topic variety within multiple query documents. While El-Arini and Guestrin [19] evaluates their approach with a user study, Jacovi et al. [31] employ document key-phrases as a proxy for finer grained relevance.

Literature search and recommendation: Other related work to the one presented here comes from a range of work exploring literature search. Chakraborty et al. [12] trained faceted paper retrieval based on citation contexts in a paper. Jain et al. [32] train models to learn disentangled abstract representations trained on aspect-labelled data for biomedical randomized control trial papers. Chan et al. [13] also presents closely related work, where they explore the problem of recommending analogically similar scientific articles, and wherein they also include a small-scale evaluation dataset comparable to the one presented here. In similar vein Neves et al. [51], extensively evaluate a range of methods intended to extract rhetorical structure elements for faceted scientific paper search and evaluate it with a small scale dataset of biomedical publications labelled for fine grained facet similarity. Hope et al. [28] allows exploratory search using multi-facet characterizations of scientific articles for the COVID19 research literature. Faceted document similarity for articles has been explored most recently in by Ostendorff [53], Ostendorff et al. [54] and Kobayashi et al. [36]. Both Ostendorff et al. [54] and Kobayashi et al. [36] evaluate using the task of predicting facets of similarity for papers cited in a particular section of a research article. Kobayashi et al. [36] further evaluate on a context-dependent co-citation ranking task as well. Work presented in Ostendorff [53] and Kobayashi et al. [36] present tasks most similar to ours while lacking in manually annotated datasets.

A range of work also explores the problem of document representations for an unfaceted content based recommendations in scientific documents [16, 8] and are often evaluated using citation or paper recommendation tasks – Färber and Jatowt [21] provide an extensive survey of this literature.

Other Datasets: A handful of other work also bears resemblance to aspects of our dataset. Brown et al. [10], present a large expert annotated biomedical dataset intended to benchmark content based literature recommendation, however the dataset is annotated at the whole abstract level unlike faceted relevances as presented here. Datasets intended to match citation context sentences to the abstract sentences of the papers they cite share some similarity with the task of querying with sentences of a facet [14]. However they represent a somewhat simpler and different task, requiring to match sentences from full-text to the citance. In similar vein, our task setup also resembles that of claim verification as proposed in Wadden et al. [68]. This setup however while serving the different goal of claim verifications also deals with atomic facts as opposed to more complex context dependent scientific paper facets as in our setup. Finally, datasets intended for citation intent prediction [33, 50, 63] mainly focus on a classification task involving predicting citation-intent given pairs of papers as opposed to retrieving papers while conditioning on a paper and a facet.

8 Conclusions

In this work we formalize the task of faceted Query by Example in the context of scientific literature search and highlight several important related problems our dataset can facilitate progress on. Given the problems of scientific search, the inability to articulate keyword queries, context dependence, and desire for exploratory search, we believe the faceted QBE formulation provides meaningful benefit. We provide a expert annotated test collection for the evaluation of the faceted QBE task. While prior work on the problem has often relied upon small evaluation sets, silver evaluations based on citations or keyword based relevance, or expensive human evaluation we believe our dataset provides a pragmatic alternative which will facilitate comparison and development of models. Finally we evaluate performance with a host of strong baseline approaches and highlight the challenging aspects of the dataset and the faceted QBE problem in general, and note that the dataset offers significant room for improvement.

9 Acknowledgements

We are grateful to the members of IESL at UMass Amherst and the Olivetti Group at MIT for helpful discussions in formulating this project. We also appreciate the contributions of our annotators in creating the annotation guidelines and the dataset. We also thank anonymous reviewers for their suggestions for making our work stronger. This work was supported in part by the National Science Foundation under Grant No. DMR-1922090, the Chan Zuckerberg Initiative under the project Scientific Knowledge Base Construction, the International Business Machines Corporation Cognitive Horizons Network under the project Knowledge Extraction, Representation, and Reasoning, and the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

References

[1] Yalemisew Abgaz, Diarmuid O'Donoghue, Dmitry Smorodinnikov, and Donny Hurley. 2016. Evaluation often Analogical Inferences Formed from Automatically Generated Representations of Scientific Publications. (2016).

[2] Tobi Adewoye, Xiao Han, Nick Ruest, Ian Milligan, Samantha Fritz, and Jimmy Lin. 2020. Content-Based Exploration of Archival Images Using Neural Networks. ACM/IEEE. `https://dl.acm.org/doi/10.1145/3383583.3398577`

[3] Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez Agirre, Rada Mihalcea, German Rigau Claramunt, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval-2016*. ACL.

[4] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying Search Results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining* (Barcelona, Spain) *(WSDM '09)*. Association for Computing Machinery, New York, NY, USA, 5–14. `https://doi.org/10.1145/1498759.1498766`

[5] Richard Antonello, Nicole Beckage, Javier Turek, and Alexander Huth. 2021. Selecting Informative Contexts Improves Language Model Fine-tuning. In *ACL*. Association for Computational Linguistics, Online. `https://aclanthology.org/2021.acl-long.87`

[6] Jaime Arguello, Adam Ferguson, Emery Fine, Bhaskar Mitra, Hamed Zamani, and Fernando Diaz. 2021. Tip of the Tongue Known-Item Retrieval: A Case Study in Movie Identification. In *Proceedings of the 6th international ACM SIGIR Conference on Human Information Interaction and Retrieval*. ACM. `https://dlnext.acm.org/doi/10.1145/3406522.3446021`

[7] Mark Berger, Jakub Zavrel, and Paul Groth. 2020. Effective distributed representations for academic expert search. In *Proceedings of the First Workshop on Scholarly Document Processing*. Association for Computational Linguistics, Online, 56–71. `https://doi.org/10.18653/v1/2020.sdp-1.7`

[8] 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)*. Association for Computational Linguistics, New Orleans, Louisiana, 238–251. `https://doi.org/10.18653/v1/N18-1022`

[9] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 632–642. `https://doi.org/10.18653/v1/D15-1075`

[10] Peter Brown, RELISH Consortium, and Yaoqi Zhou. 2019. Large expert-curated database for benchmarking document similarity detection in biomedical literature search. *Database* 2019 (10 2019). `https://doi.org/10.1093/database/baz085`

[11] Hancheng Cao, Mengjie Cheng, Zhepeng Cen, Daniel McFarland, and Xiang Ren. 2020. Will This Idea Spread Beyond Academia? Understanding Knowledge Transfer of Scientific Concepts across Text Corpora. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 1746–1757. `https://doi.org/10.18653/v1/2020.findings-emnlp.158`

[12] Tanmoy Chakraborty, Amrith Krishna, Mayank Singh, Niloy Ganguly, Pawan Goyal, and Animesh Mukherjee. 2016. FeRoSA: A Faceted Recommendation System for Scientific Articles. In *Proceedings, Part II, of the 20th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining - Volume 9652* (Auckland, New Zealand) *(PAKDD 2016)*. Springer-Verlag, Berlin, Heidelberg, 528–541. https://doi.org/10.1007/978-3-319-31750-2_42

[13] Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. 2018. SOLVENT: A Mixed Initiative System for Finding Analogies between Research Papers. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 31 (Nov. 2018), 21 pages. https://doi.org/10.1145/3274300

[14] Muthu Kumar Chandrasekaran, Michihiro Yasunaga, Dragomir Radev, Dayne Freitag, and Min-Yen Kan. 2019. Overview and Results: CL-SciSumm Shared Task 2019. In *In Proceedings of Joint Workshop on Bibliometric-enhanced Information Retrieval and NLP for Digital Libraries (BIRNDL 2019)*. http://ceur-ws.org/Vol-2414/paper17.pdf

[15] Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019. Pretrained Language Models for Sequential Sentence Classification. 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)*. Association for Computational Linguistics, Hong Kong, China, 3693–3699. https://doi.org/10.18653/v1/D19-1383

[16] 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*. Association for Computational Linguistics, Online, 2270–2282. https://doi.org/10.18653/v1/2020.acl-main.207

[17] Kyriaki Dimitriadou, Olga Papaemmanouil, and Yanlei Diao. 2014. Explore-by-Example: An Automatic Query Steering Framework for Interactive Data Exploration. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data* (Snowbird, Utah, USA) *(SIGMOD '14)*. Association for Computing Machinery, New York, NY, USA, 517–528. https://doi.org/10.1145/2588555.2610523

[18] Cody Dunne, Ben Shneiderman, Robert Gove, Judith Klavans, and Bonnie Dorr. 2012. Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization. *Journal of the American Society for Information Science and Technology* 63, 12 (2012), 2351–2369. https://doi.org/10.1002/asi.22652 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.22652

[19] Khalid El-Arini and Carlos Guestrin. 2011. Beyond Keyword Search: Discovering Relevant Scientific Literature. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Diego, California, USA) *(KDD '11)*. Association for Computing Machinery, New York, NY, USA, 439–447. https://doi.org/10.1145/2020408.2020479

[20] Marzieh Fadaee, Olga Gureenkova, Fernando Rejon Barrera, Carsten Schnober, Wouter Weerkamp, and Jakub Zavrel. 2020. A New Neural Search and Insights Platform for Navigating and Organizing AI Research. In *Proceedings of the First Workshop on Scholarly Document Processing*. Association for Computational Linguistics, Online, 207–213. https://doi.org/10.18653/v1/2020.sdp-1.23

[21] Michael Färber and Adam Jatowt. 2020. Citation recommendation: approaches and datasets. *International Journal on Digital Libraries* 21, 4 (2020), 375–405. https://doi.org/10.1007/s00799-020-00288-2

[22] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 6894–6910. https://aclanthology.org/2021.emnlp-main.552

[23] Dedre Gentner and Kenneth D. Forbus. 2011. Computational models of analogy. *WIREs Cognitive Science* 2, 3 (2011), 266–276. https://doi.org/10.1002/wcs.105 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcs.105

[24] Yolanda Gil, Mark Greaves, James Hendler, and Haym Hirsh. 2014. Amplify scientific discovery with artificial intelligence. *Science* 346, 6206 (2014), 171–172. `https://doi.org/10.1126/science.1259439` arXiv:https://science.sciencemag.org/content/346/6206/171.full.pdf

[25] Daniel S. Hain, Roman Jurowetzki, Tobias Buchmann, and Patrick Wolf. 2020. Text-based Technological Signatures and Similarities: How to create them and what to do with them. *ArXiv* abs/2003.12303 (2020).

[26] Yun He, Zhuoer Wang, Yin Zhang, Ruihong Huang, and James Caverlee. 2020. PARADE: A New Dataset for Paraphrase Identification Requiring Computer Science Domain Knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 7572–7582. `https://doi.org/10.18653/v1/2020.emnlp-main.611`

[27] Lauri Himanen, Amber Geurts, Adam Stuart Foster, and Patrick Rinke. 2019. Data-Driven Materials Science: Status, Challenges, and Perspectives. *Advanced Science* 6, 21 (2019), 1900808. `https://doi.org/10.1002/advs.201900808` arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/advs.201900808

[28] Tom Hope, Jason Portenoy, Kishore Vasan, Jonathan Borchardt, Eric Horvitz, Daniel S Weld, Marti A Hearst, and Jevin West. 2020. SciSight: Combining faceted navigation and research group detection for COVID-19 exploratory scientific search. *arXiv preprint arXiv:2005.12668* (2020).

[29] Tom Hope, Ronen Tamari, Hyeonsu Kang, Daniel Hershcovich, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2021. Scaling Creative Inspiration with Fine-Grained Functional Facets of Product Ideas. arXiv:2102.09761

[30] Chien-yu Huang, Arlene Casey, Dorota Głowacka, and Alan Medlar. 2019. Holes in the Outline: Subject-Dependent Abstract Quality and Its Implications for Scientific Literature Search *(CHIIR '19)*. Association for Computing Machinery, New York, NY, USA, 289–293. `https://doi.org/10.1145/3295750.3298953`

[31] Alon Jacovi, Gang Niu, Yoav Goldberg, and Masashi Sugiyama. 2021. Scalable Evaluation and Improvement of Document Set Expansion via Neural Positive-Unlabeled Learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 581–592. `https://www.aclweb.org/anthology/2021.eacl-main.47`

[32] Sarthak Jain, Edward Banner, Jan-Willem van de Meent, Iain J. Marshall, and Byron C. Wallace. 2018. Learning Disentangled Representations of Texts with Application to Biomedical Abstracts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 4683–4693. `https://doi.org/10.18653/v1/D18-1497`

[33] 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 (2018), 391–406. `https://doi.org/10.1162/tacl_a_00028`

[34] Herman Kamper, Aristotelis Anastassiou, and Karen Livescu. 2019. Semantic Query-by-example Speech Search Using Visual Grounding. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7120–7124. `https://doi.org/10.1109/ICASSP.2019.8683275`

[35] Maryam Karimzadehgan, ChengXiang Zhai, and Geneva Belford. 2008. Multi-Aspect Expertise Matching for Review Assignment. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management* (Napa Valley, California, USA) *(CIKM '08)*. Association for Computing Machinery, New York, NY, USA, 1113–1122. `https://doi.org/10.1145/1458082.1458230`

[36] Yuta Kobayashi, Masashi Shimbo, and Yuji Matsumoto. 2018. Citation Recommendation Using Distributed Representation of Discourse Facets in Scientific Articles. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries* (Fort Worth, Texas, USA) *(JCDL '18)*. Association for Computing Machinery, New York, NY, USA, 243–251. `https://doi.org/10.1145/3197026.3197059`

[37] Weize Kong. 2016. Extending Faceted Search to the Open-Domain Web. *SIGIR Forum* 50, 1 (June 2016), 90–91. `https://doi.org/10.1145/2964797.2964814`

[38] Weize Kong. 2016. Extending Faceted Search to the Open-Domain Web. *SIGIR Forum* 50, 1 (June 2016), 90–91. `https://doi.org/10.1145/2964797.2964814`

[39] Weize Kong and James Allan. 2016. Precision-Oriented Query Facet Extraction. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (Indianapolis, Indiana, USA) *(CIKM '16)*. Association for Computing Machinery, New York, NY, USA, 1433–1442. `https://doi.org/10.1145/2983323.2983824`

[40] Ruben Kruiper, Jessica Chen-Burger, and Marc P. Y. Desmulliez. 2016. Computer-Aided Biomimetics. In *Biomimetic and Biohybrid Systems*, Nathan F. Lepora, Anna Mura, Michael Mangan, Paul F.M.J. Verschure, Marc Desmulliez, and Tony J. Prescott (Eds.). Springer International Publishing, Cham, 131–143. `https://doi.org/10.1007/978-3-319-42417-0_13`

[41] Alex Ksikes. 2014. *Towards exploratory faceted search systems*. Ph.D. Dissertation. University of Cambridge. `https://doi.org/10.17863/CAM.14080`

[42] Sara Katherine Lafia. 2020. *Designing for Serendipity: Research Data Curation in Topic Spaces*. Ph.D. Dissertation. UC Santa Barbara. `https://escholarship.org/uc/item/5647q82f`

[43] Esther Landhuis. 2016. Scientific literature: Information overload. *Nature* 535, 7612 (2016), 457–458. `https://doi.org/10.1038/nj7612-457a`

[44] N. Lavrac, Matej Martinc, Senja Pollak, and B. Cestnik. 2020. Bisociative Literature-Based Discovery: Lessons Learned and New Prospects. In *ICCC*. `http://computationalcreativity.net/iccc20/papers/034-iccc20.pdf`

[45] Matteo Lissandrini, Davide Mottin, Themis Palpanas, and Yannis Velegrakis. 2019. Example-Based Search: A New Frontier for Exploratory Search. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 1411–1412. `https://doi.org/10.1145/3331184.3331387`

[46] 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*. Association for Computational Linguistics, Online, 4969–4983. `https://doi.org/10.18653/v1/2020.acl-main.447`

[47] Douglas L. Medin, Robert L. Goldstone, and Dedre Gentner. 1990. Similarity Involving Attributes and Relations: Judgments of Similarity and Difference Are Not Inverses. *Psychological Science* 1, 1 (1990), 64–69. `http://www.jstor.org/stable/40062393`

[48] Reagan Mozer, Luke Miratrix, Aaron Russell Kaufman, and L. Jason Anastasopoulos. 2020. Matching with Text Data: An Experimental Evaluation of Methods for Matching Documents and of Measuring Match Quality. *Political Analysis* 28, 4 (2020), 445–468. `https://doi.org/10.1017/pan.2020.1`

[49] Ramesh Nallapati, Sanga Peerreddy, and Prateek Singhal. 2012. *Skierarchy: Extending the power of crowdsourcing using a hierarchy of domain experts, crowd and machine learning*. Technical Report. `https://apps.dtic.mil/sti/citations/ADA581773`

[50] Hidetsugu Nanba and Manabu Okumura. 1999. Towards Multi-Paper Summarization Reference Information. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2* (Stockholm, Sweden) *(IJCAI'99)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 926–931.

[51] Mariana Neves, Daniel Butzke, and Barbara Grune. 2019. Evaluation of Scientific Elements for Text Similarity in Biomedical Publications. In *Proceedings of the 6th Workshop on Argument Mining*. Association for Computational Linguistics, Florence, Italy. `https://www.aclweb.org/anthology/W19-4515`

[52] OpenReview. [n.d.]. Paper-reviewer affinity modeling for OpenReview. `https://github.com/openreview/openreview-expertise`. Accessed: 27 August 2021.

[53] Malte Ostendorff. 2020. Contextual Document Similarity for Content-based Literature Recommender Systems. *Proceedings of the Doctoral Consortium at ACM/IEEE Joint Conference on Digital Libraries*.

[54] Malte Ostendorff, Terry Ruas, Till Blume, Bela Gipp, and Georg Rehm. 2020. Aspect-based Document Similarity for Research Papers. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 6194–6206. https://doi.org/10.18653/v1/2020.coling-main.545

[55] Elisabeth Pain. 2016. How to keep up with the scientific literature. *Science Careers* 30 (2016). https://www.science.org/content/article/how-keep-scientific-literature-rev2

[56] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[57] Margaret E. Roberts, Brandon M. Stewart, and Richard A. Nielsen. 2020. Adjusting for Confounding with Text Matching. *American Journal of Political Science* 64, 4 (2020), 887–903. https://doi.org/10.1111/ajps.12526

[58] Sheikh Muhammad Sarwar and James Allan. 2020. Query by Example for Cross-Lingual Event Retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 1601–1604. https://doi.org/10.1145/3397271.3401283

[59] Sheikh Muhammad Sarwar and James Allan. 2020. Query by Example for Cross-Lingual Event Retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 1601–1604. https://doi.org/10.1145/3397271.3401283

[60] Semantic Scholar. [n.d.]. Semantic Scholar on Twitter. https://twitter.com/SemanticScholar/status/1267867735318355968/retweets. Accessed: 27 August 2021.

[61] Micah Shlain, Hillel Taub-Tabib, Shoval Sadde, and Yoav Goldberg. 2020. Syntactic Search by Example. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Online, 17–23. https://doi.org/10.18653/v1/2020.acl-demos.3

[62] Hillel Taub Tabib, Micah Shlain, Shoval Sadde, Dan Lahav, Matan Eyal, Yaara Cohen, and Yoav Goldberg. 2020. Interactive Extractive Search over Biomedical Corpora. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*. Association for Computational Linguistics, Online, 28–37. https://doi.org/10.18653/v1/2020.bionlp-1.3

[63] 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*. Association for Computational Linguistics, Sydney, Australia, 103–110. https://aclanthology.org/W06-1613

[64] Daniel Tunkelang. 2006. Dynamic category sets: An approach for faceted search. In *ACM SIGIR*, Vol. 6. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.83.4202&rep=rep1&type=pdf

[65] Prajna Upadhyay, Srikanta Bedathur, Tanmoy Chakraborty, and Maya Ramanath. 2020. Aspect-Based Academic Search Using Domain-Specific KB. In *Advances in Information Retrieval*, Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins (Eds.). Springer International Publishing, Cham, 418–424. https://link.springer.com/chapter/10.1007/978-3-030-45442-5_52

[66] Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. 2014. OpenML: Networked Science in Machine Learning. *SIGKDD Explor. Newsl.* 15, 2 (June 2014), 49–60. https://doi.org/10.1145/2641190.2641198

[67] Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. TREC-COVID: Constructing a Pandemic Information Retrieval Test Collection. *SIGIR Forum*, Article 1 (Feb. 2021), 12 pages. https://doi.org/10.1145/3451964.3451965

[68] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 7534–7550. https://doi.org/10.18653/v1/2020.emnlp-main.609

[69] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. 2013. A Theoretical Analysis of NDCG Type Ranking Measures. In *Proceedings of the 26th Annual Conference on Learning Theory (Proceedings of Machine Learning Research, Vol. 30)*, Shai Shalev-Shwartz and Ingo Steinwart (Eds.). PMLR, Princeton, NJ, USA, 25–54. https://proceedings.mlr.press/v30/Wang13.html

[70] Mingzhu Zhu and Yi-Fang Brook Wu. 2014. Search by Multiple Examples. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining* (New York, New York, USA) *(WSDM '14)*. Association for Computing Machinery, New York, NY, USA, 667–672. https://doi.org/10.1145/2556195.2556206

A Checklist Materials

Datasheet: https://github.com/iesl/CSFCube/blob/master/datasheet.md

License: https://github.com/iesl/CSFCube/blob/master/LICENSE.md

B Full Text vs Abstract Annotations

As we note in §4, to examine the effect of annotating the abstract of papers instead of full-text of papers we also conduct a small scale study to examine the scale of differences between the relevance ratings produced each of them. This is done by a single expert annotator annotating relevance based on abstract and full-text separately for 9 query abstracts (3 from each facet) and 5 candidate abstracts each (45 pairs). In making these annotations, queries were picked to ensure all paper-types (§3) were represented and candidates were chosen at random from across all relevance levels. Next abstract based relevances were labelled, following this full-text relevances were labelled. In labelling full-text relevances care was taken to not show abstract based relevances or the abstract text. In making full-text judgements the paper was skimmed for content relevant to the target facet rather than read exhaustively. Every full-text judgement pair took about 5 minutes to complete. Finally, while the presented study isnt intended to be statistically robust we believe it presents a reasonable pilot study in support of the abstract based ratings adopted in our dataset annotation.

C Baseline Methods

The methods we choose to evaluate capture a range of granularities and nature of methods: term based methods, pre-trained model based sentence representations, and whole abstract representation models. Note that some of the methods evaluated are included in our set of methods to construct candidate pools, but as noted in §3 they used unfaceted representations.

fabs_tfidf: This is a simple faceted baseline which builds a sparse TF-IDF representation for the sentences corresponding to the query facet in the query abstract. Candidates are represented by their whole abstract TF-IDF representations.

fabs_bm25: This represents a baseline identical to fabs_tfidf while using the Okapi BM25 weighting scheme.[8]

fabs_cbow200: This is a dense bag-of-words representation for the sentences corresponding to the query facet in the query abstract – token embeddings are averaged. As above, candidates are represented by all abstract sentences. The word2vec embeddings are trained on $800,000$ abstracts from the S2ORC corpus (§3). We used 200 dimensional word embeddings.

fabs_tfidfcbow200: This baseline combines the above baselines where the word2vec representations are weighted by TF-IDF weights prior to being averaged.

SentBERT: SentBERT [56] represents a sentence level model. In our setup we encode all query facet sentences and all candidate abstract sentences individually with SentBERT, and then use the maximum pairwise sentence cosine similarity between the query and candidate sentences to rank candidates. We evaluate two versions of SentBERT, one fine-tuned only on Natural Language Inference (NLI) datasets as in Reimers and Gurevych [56] and a second model fine-tuned on NLI and a wide variety of paraphrase text. We term these SentBERT-NLI and SentBERT-PP.[9] We choose to use SentBERT-PP given its strong performance on the SciDOCS benchmark [16].[10]

SimCSE: SimCSE [22], represents a very recent state of the art sentence similarity model trained in two ways – an unsupervised manner training an encoder to maximise similarity with a a "dropped-out" representation of the same sentence, and a supervised version trained on NLI data. We denote these as UnSimCSE and SuSimCSE. We use the models princeton-nlp/unsup-simcse-bert-base-uncased and princeton-nlp/sup-simcse-bert-base-uncased made available through the Hugging Face[11] package.

[8] BM25 implementation: https://github.com/dorianbrown/rank_bm25

[9] we use the sentence_transformers package. In this package, SentBERT-NLI corresponds to nli-roberta-base-v2 and SentBERT-PP to paraphrase-TinyBERT-L6-v2.

[10] Model performances: https://www.sbert.net/docs/pretrained_models.html

[11] https://huggingface.co/

SPECTER: This approach represents a multi-layer transformer based SciBERT model fine-tuned on citation network data [16]. SPECTER operates on titles and the whole abstract of the papers and represents an entirely un-faceted model. Both queries and candidates are represented by their SPECTER embeddings. Note that SPECTER was trained on a corpus of randomly selected scientific documents. We also re-implement and train a version of SPECTER on about 660k computer science paper triples with identical hyper-parameters to SPECTER, we call this in-domain model SPECTER-ID.

In re-ranking the candidate pool for every query, the L2 distance between the query and candidate vectors was used unless specified otherwise.

D Evaluating Citations

Because we included cited papers in our candidate pools, and manually assign relevance judgments for them, this dataset allows us to examine the common assumption that cited paper abstracts will be relevant to a query paper abstract. In this analysis, we find cited papers to pre-dominantly be rated at 0 or 1 levels of relevance, 79% of the times for `background`, 88% of the times for `method`, and 86% of the times for `result`. Given that citations are often considered incidental signals from which to train models and often to evaluate them as well, we believe these observations will have implications for future modeling and evaluation work. We hope future work will use citations with caution, particularly in evaluation setups for tasks similar to the one posed here.

E Potential Applications

A range of important applications rely on computing similarity between scientific texts. Given that our dataset allow evaluation of document similarity methods in general we believe our test collection fills an important gap in the development and benchmarking of methods intended for these applications.

Exploratory Search: Content based search with paradigms such as Query by Example has been considered more suited to exploratory search tasks [41, 17, 45] than keyword based search. Recent work has also seen AI powered literature navigation tools leveraging content based search at varying levels of granularity [20]. We believe our task formulation directly suits this and will allow development of methods intended for these applications.

Patent Search: Hain et al. [25], highlight the case of measuring technological similarities between patents based on the abstracts of patents, and the subsequent employment of this information in mapping patent quality and mapping technological change. Further they also highlight the lack of benchmarks for the measurement of technological similarity between patents [25, Sec 4.3]. While differing in domain we believe our work provides a valuable resource for model development.

Text Matching for Causal Inference: Mozer et al. [48] highlight the importance of text matching for causal inference from observational text data: "matched documents can be used to make unbiased comparisons between groups on external features such as rates of citation". Roberts et al. [57], demonstrate just such a investigation into the gendered biases of citation patterns. The reliance of these analysis on document similarity and matching across a corpus along specific aspects allows our dataset to be of value in developing methods for document matching.

Expert Search: Expert search presents an important application, specially in the contex of peer review, where scientific papers must be matched to experts suited to review it. This often involves computing scientific document similarity [7], a venue where our work proves valuable. In the case of work such as Karimzadehgan et al. [35], which attempts to find experts along all aspects of a scientific paper, our work provides an even stronger resource.

F Extended results

While Section 5 presents NDCG$_{\%20}$, we additionally report NDCG$_{\%100}$ in extended results. NDCG$_{\%100}$ indicates a metric comuted based on the entire pool per query. We note based on Wang et al. [69], that larger pools cause larger values of NDCG, this is observed here. Further model performance at lower values of k, i.e. at the top of the predicted rankings, is still lower indicating

Table 4: Extended test set results for the set of baselines methods. Metrics (R-Precision, Precision and Recall at 20, NDCG$_{\%20}$, NDCG$_{\%100}$) are computed based on a 2-fold cross-validation, represent averages over per-query metrics, and are reported as percentages. SPECTER-ID performance is reported over three training re-runs, the remaining baselines are reported based on a single set of model parameters released by the respective authors.

	`background`					`method`				
	RP	P@20	R@20	NDCG$_{\%100}$	NDCG$_{\%20}$	RP	P@20	R@20	NDCG$_{\%100}$	NDCG$_{\%20}$
`fabs_tfidf`	23.35	27.19	45.80	78.14	57.97	09.30	09.83	34.75	57.87	31.20
`fabs_bm25`	20.12	27.81	49.85	79.02	59.39	09.37	11.63	38.29	60.68	34.59
`fabs_cbow200`	19.61	15.94	27.97	67.68	36.56	08.65	08.33	15.69	51.58	21.14
`fabs_tfidfcbow200`	15.92	16.87	27.76	69.77	40.51	07.99	06.01	17.71	51.87	21.70
`SentBERT-PP`	21.24	28.75	46.67	79.14	60.80	10.00	10.83	36.30	59.50	33.40
`SentBERT-NLI`	19.02	25.00	40.13	75.80	54.23	09.11	11.46	02.89	58.52	31.10
`UnSimCSE-BERT`	18.15	23.44	36.05	74.34	51.59	08.86	09.65	27.92	59.21	31.23
`SuSimCSE-BERT`	19.22	22.81	46.75	76.70	55.22	08.58	09.76	29.01	58.54	30.88
SPECTER	**24.81**	**35.31**	**57.45**	82.24	66.70	**11.72**	13.58	40.81	62.77	37.41
SPECTER-ID	24.55 ±1.3	34.17 ±0.5	53.26 ±0.3	**84.31** ±0.8	**69.22** ±1.71	10.53 ±0.3	**16.22** ±1.21	**44.59** ±3.6	**64.63** ±0.4	**42.76** ±0.78

	`result`					`all`				
	RP	P@20	R@20	NDCG$_{\%100}$	NDCG$_{\%20}$	RP	P@20	R@20	NDCG$_{\%100}$	NDCG$_{\%20}$
`fabs_tfidf`	11.35	16.28	38.57	66.12	41.24	14.59	17.64	39.69	67.17	43.19
`fabs_bm25`	11.31	20.00	40.40	67.87	45.07	13.50	19.69	42.73	68.97	46.06
`fabs_cbow200`	11.16	10.42	23.44	60.22	30.93	13.08	11.47	22.23	59.64	29.36
`fabs_tfidfcbow200`	10.43	10.69	24.39	60.30	32.79	11.38	11.09	23.13	60.42	31.42
`SentBERT-PP`	13.60	19.83	41.73	71.90	52.35	14.83	19.62	41.41	69.98	48.57
`SentBERT-NLI`	14.23	22.05	46.99	72.13	51.30	14.04	19.42	38.67	68.68	45.39
`UnSimCSE-BERT`	12.00	19.58	38.95	68.44	45.55	12.92	17.41	34.43	67.17	42.59
`SuSimCSE-BERT`	12.37	18.58	39.76	68.78	44.93	13.33	16.95	34.83	67.83	43.45
SPECTER	18.62	23.78	52.72	75.47	56.67	18.29	23.97	50.14	73.30	53.28
SPECTER-ID	**20.09** ±0.92	**27.36** ±0.45	**58.74** ±3.04	**76.49** ±0.41	**60.40** ±1.31	**18.32** ±0.79	**25.74** ±0.22	**52.12** ±1.54	**74.96** ±0.06	**57.22** ±0.70

significant room for improvements. Finally, note that an apt value of k in computing metrics for evaluation will depend on the choice of target application, we believe trends highlighted in our results hold across values of k as per the consistency of relative performance of models on NDCG$_{\%20}$ and NDCG$_{\%100}$.

G Error Analysis

Based on a qualitative examination of per-query ranking performance of `abs_tfidf`, `SentBERT-PP` and SPECTER-ID we outline a range of factors which lead the baseline models to underperform. We believe the incorporation of modeling to handle these phenomena will lead to improved performance on our dataset. We indicate various error cases through examples of the query facet, false positive top retrievals (FP), or false negative lower ranked retrievals (FN). We mention the query ID for examples in superscripts, use underlines to emphasize important segments, and we only provide the relevant sentences from the abstract in each example due to space constraints.

Salient Aspects: One source of error is the inability of models to identify the most salient aspects for similarity, often expressed only in part of a larger set of facet sentences.

`background Q:` “Many classification problems require decisions among a large number of competing classes.”^{1791179}

FP: “Several real problems involve the classification of data into categories or classes.”

`background Q:` “With the increasing empirical success of distributional models of compositional semantics, it is timely to consider the types of textual logic that such models are capable of capturing. In this paper, we address shortcomings in the ability of current models to capture logical operations such as negation.”^{1936997}

Nearly all models miss the notion of negation in the above example.

Multiple Aspects: Within a given facet, papers often expressed multiple finer grained aspects, models however often only retrieved based on a single aspect. In the following baseline models often retrieved based on one or the other aspect:

`method Q:` “We present a Few-Shot Relation Classification Dataset (FewRel), . . . The relation of each sentence is first recognized by distant supervision methods, and then filtered by crowdworkers. We adapt the most recent state-of-the-art few-shot learning methods for relation classification and conduct a thorough evaluation of these methods."[53080736]

Domain specific similarities A set of errors also arise from the inability of models to determine similarity between technical concepts. The example represents an inability to rate "stacking", "ensemble strategy", and "bagging" as similar.

`result Q:` "Using a public corpus, we show that stacking can improve the efficiency of automatically induced anti-spam filters, . . . "[3264891]

FN: "The experiments on standard WEBSPAM-UK2006 benchmark showed that the ensemble strategy can improve the web spam detection performance effectively."

FN: "We evaluate the classifier performances and find that BAGGING performs the best. . . . our method may be an excellent means to classify spam emails"

Mechanistic similarities: Nearly all methods perform poorly in the case of determining mechanistic similarity in `method` facets. This often relies on determining similarity across a sequence of actions. Baseline models failed to align steps [1] and [2] across abstracts below.

`method Q:` "Using an annotated set of"factual"and"feeling"debate forum posts, [1]we extract patterns that are highly correlated with factual and emotional arguments, and [2]then apply a bootstrapping methodology to find new patterns in a larger pool of unannotated forum posts.[10010426]"

FN: "[1]High-precision classifiers label unannotated data to automatically create a large training set, which is then given to an extraction pattern learning algorithm. [2]The learned patterns are then used to identify more subjective sentences."

Context dependence of facets: Faceted similarities as labelled here often also show context dependence on other facets. This is notable in the case of `result` queries. Given that one major guideline for result similarity in our dataset are if "the same finding or conclusion" is found, being able to determine context similarity is important.

`result Q:` ". . . Subsequently, lexical cue proportions, predicted certainty, as well as their time course characteristics are used to compute veracity for each rumor tweet . . . . Evaluated on the data portion for which hand-labeled examples were available, it achieves .74 F1-score on identifying rumor resolving tweets and .76 F1-score on predicting if a rumor is resolved as true or false.[5052952]

FN: "In this study, we propose a novel approach to capture the temporal characteristics of these features based on the time series of rumor's lifecycle, for which time series modeling technique is applied to incorporate various social context information. Our experiments using the events in two microblog datasets confirm that the method outperforms state-of-the-art rumor detection approaches by large margins."

In these examples, determining that the higher level result of time series information being important for identifying rumour tweets relies on modeling method similarity. We believe approaches which improve upon `method` similarity, will likely benefit overall performance on other facets as well.

Qualitative result statements: Finally, we also note that `result` queries which summarize qualitative findings often perform poorer, often requiring broader context and often lacking in term overlaps which may otherwise easily indicate relevance.

`result Q:` "Experiments with several Reddit forums show that style is a better indicator of community identity than topic, even for communities organized around specific topics. Further, there is a positive correlation between the community reception to a contribution and the style similarity to that community, but not so for topic similarity."[11629674]

H Potential training data sources

Given these challenges, we also highlight specific other sources of data that future work may exploit to train models to overcome these problems:

Domain specific paraphrase datasets: Given the reasonably strong performance of the `SentBERT-PP` model, fine-tuned on paraphrase datasets, we believe other domain specific paraphrase datasets have the potential to be useful for the proposed task. An example is PARADE [26] which presents a dataset of computer science paraphrase pairs.

Selecting informative citation examples: Appendix D presents an analysis of citation data and indicates how only a part of this data contains fine-grained facet similarities. An potential approach to selecting more informative citation examples might involve model dependent training data subset selection approaches such as that proposed in Antonello et al. [5].

Co-citations data: Given that the proposed task relies on capturing fine-grained similarities, co-citations examples in the full-text of papers – papers cited in a narrow context (such as a sentence or paragraph), also promise to contain finer grained similarities likely to help train better models [36]. Use of these examples is specially promoted by existence of parsed full-text data in in the S2ORC corpus.