Multi-Label Field Classification for Scientific Documents using Expert and Crowd-sourced Knowledge

Rebecca Gelles
CSET, Georgetown University
500 1st St NW, Washington DC 20001
rebecca.gelles@georgetown.edu

James Dunham
CSET, Georgetown University
500 1st St NW, Washington DC 20001
james.dunham@georgetown.edu

Abstract

Taxonomies of scientific research seek to describe complex domains of activity that are overlapping and dynamic. We address this challenge by combining knowledge curated by the Wikipedia community with the input of subject-matter experts to identify, define, and validate a system of 1,110 granular fields of study for use in multi-label classification of scientific publications. The result is capable of categorizing research across subfields of artificial intelligence, computer security, semiconductors, genetics, virology, immunology, neuroscience, biotechnology, and bioinformatics. We then develop and evaluate a solution for zero-shot classification of publications in terms of these fields.

1 Introduction

Organizing and categorizing scholarly literature is a salient challenge for researchers, funding organizations, and data providers. Developing a comprehensive yet efficient classification system that captures the breadth and depth of the scholarly literature is a problem that has both captivated and vexed researchers. Thorough taxonomic assignment of fields would provide great value via searching and indexing capabilities, for use in research, policy, and the public good. But manual categorization is slow, expensive, and can be error-prone. The cost of manual assignment also scales with the number of fields assigned; the more comprehensive the solution, the more difficult it is for annotators to apply it. Without automation, ideally with a technique efficient and affordable enough to handle the constantly-increasing flow of scholarly data available, a broadly-usable solution will never be realistic.

This doesn't mean there is no place for manual or tailored solutions within the space of topical classification; however, manual work is best used in tandem with automated solutions. In our paper, we introduce a solution that begins with the curation of custom field taxonomies, developed with the aid of Wikipedia, existing academic taxonomies, and subject matter experts. The result is a field of study model that creates automated zero-shot field relevance scores based on Wikipedia and Wikipedia citation data. This solution combines the best of both worlds from the manual and the automated. Leveraging quality existing knowledge bases like Wikipedia ensures that fields are clearly defined and fit into well-organized hierarchies, while use of embeddings and similarity scores to produce final results allows the actual training and labeling process, the most expensive component, to be fast and automatic.

Our methodology takes the Wikipedia text of our chosen fields and the text of the page's citations and represents it in embedding form; using these embeddings allows us to compute cosine similarities between the resultant embedding and the text embedding of any given publication, creating field scores. This allows us to determine which fields are the most similar to any given publication. This methodology is fast and affordable, as we are using low-cost embedding methods, and cosine similarity is easy to calculate. Our field definitions are also highly extensible. We use a slightly modified version of Shen et al. (2018)'s field hierarchy for the top two levels (L0 and L1) of our hierarchy, adding and cleaning up fields through manual review. This yields a set of fields that are broad and complete at higher levels, and cover the full scope of the scientific literature. However, at the lower two levels (L2 and L3), we focus on specific research areas of particular interest to us, curating our subfields with support from existing academic taxonomies, Wikipedia, and subject matter experts. This means that anyone with interest in particular research areas could define their own taxonomies following the same process and use the identical method to produce field embeddings and scores for their own subfields of interest.

Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia, pages 14–20
November 16, 2024 ©2024 Association for Computational Linguistics

Another advantage of our methodology is the use of multi-label classifications. Most scientific research publications may not naturally fall into only one field, but will instead be relevant to multiple areas; this is particularly true as the relevant fields become more granular. Multi-label classifications accommodate this nuance, while the inclusion of scores allows us to step back and limit to top fields where that is preferred, or set our threshold of similarity at any given point of interest.

As our technique is unsupervised, and we do not have a ground-truth dataset, we instead evaluate our results through a variety of other mechanisms, including an examination of the embedding space, "silver" label matching of our fields to narrowly focused topic-specific venues, and a comparison of our results to a ground-truth dataset whose field taxonomy only partially aligns with ours.

2 Related Works

We extend a line of research on topical classification for scientific publications from Shen et al. (2018), who proposed zero-shot classification of papers with a taxonomy of over 200K fields following automatic hierarchical taxonomy construction per Sanderson and Croft's (1999) earlier work on subsumption. The authors reported cleaning up the top two levels of the taxonomy by hand based on their qualitative evaluation. The result of this work was available in Microsoft Academic Graph (Wang et al., 2020) before its shutdown at the end of 2021. Our work extends manual curation into a third and fourth level of this taxonomy, adding 813 new lower-level fields identified by SMEs.

Methodologically we follow Toney and Dunham (2022), who used Wikipedia page content and the text from pages' academic references to create field embeddings using a FastText (Bojanowski et al., 2017) model pre-trained on a corpus of scientific literature.

Other research has extended the approach developed by Shen et al. in different ways. OpenAlex (2022) adopted the full taxonomy from Shen et al., excluding fields with fewer than 500 tagged publications, and then trained a supervised model using the publications and field scores labeled by MAG for use in their publication dataset; essentially they considered the previous results from Shen et al. ground truth and trained a model to allow continued inference. A team at Semantic Scholar (MacMillan and Feldman, 2023) also developed a field classification model largely based on the taxonomy of Shen et al., with targeted additions based on user feedback, using a linear SVM running on character n-gram TF-IDF representations and trained on data selected by identifying venues likely to publish within a relatively narrow set of fields – an approach we use here for validation rather than as a training method.

The Field of Research Classification Shared Task at the Natural Scientific Language Processing Workshop 2024 (Ahmad et al., 2024b) addressed the problem of multi-label field classification with submissions evaluated against human labels. This task had a much narrower focus, as its taxonomy was focused specially on natural language processing rather than the whole of the scientific literature and gold data was available to train on and evaluate against, but the methodology used is still illustrative. The winning submission for the shared task, by the Bashyam and Krestel team, described in Ahmad et al. (2024a), as well as in their own paper (Bashyam and Krestel, 2024), treated the task as an extreme multi-label classification problem, extending the labeled data using weak supervision with a TF-IDF model, and then leveraging the larger set of weakly labeled data to fine-tune an X-transformer model. They applied hierarchical restrictions only after running the model, which is the same choice we ultimately make. We also evaluate our results against the gold dataset produced for the shared task.

3 Methodology

Our model was designed using three data sources. First, we identified hierarchical field taxonomies, starting with a base of the taxonomy developed by Shen et al. (2018) for Microsoft Academic Graph and then developing our own lower-level taxonomies using topic-specific resources, subject matter experts, and Wikipedia's own Category and List pages. We then used Wikipedia as a knowledge base from which to derive the individual fields of study and their definitions and to extract text, citations, and linkages for building model embeddings. Finally, we employed these resulting fields of study and their embeddings to classify a large corpus of academic publications drawn from a variety of datasets: Clarivate's Web of Science, Semantic Scholar, OpenAlex, The Lens, Papers with Code, and arXiv. Our corpus contains 207,231,266 publications overall.

As we developed our taxonomies, we began with a base of the high-level taxonomies curated by Microsoft Academic Graph (MAG) in their original version of the fields of study. We used their taxonomies for both our level zero (L0) fields and for the vast majority of our level one (L1) fields. The L0 and L1 fields in MAG were derived from the Science-Metrix classification scheme and refined manually by Shen et al. (2018), so they are generally of high quality, whereas the lower-level MAG fields were derived automatically, and we found them to be less intuitive. (They omitted significant areas of research and included ones that weren't clearly distinguishable from each other.) After consultation with subject matter experts, we refined some of the L1 fields to better reflect a more consensus view of how certain subject areas are organized. Otherwise we largely retained MAG's structure.

To define L2 and L3 fields, we began by focusing on a subset of fields of particular interest to us, and ones in which we had access to subject matter experts. Our methodology should translate to any similar subfields. For each subfield of interest, we identified existing taxonomies of relevance, often created by local conferences or journals for organizing their own work, or used at universities to describe course structures. We linked these taxonomies to their corresponding Wikipedia pages, and supplemented those using Wikipedia pages of relevance identified from Category and List pages about our subfields. We enlisted the aid of subject matter experts to expand on, clean up, and check the resulting fields. On occasions where a topic was of sufficient relevance but did not have a single specific Wikipedia page of its own, we identified sections of Wikipedia pages or combined multiple Wikipedia pages that could substitute. We created L2 and L3 fields beneath the following L1 fields: artificial intelligence, computer security, semiconductors, genetics, virology, immunology, neuroscience, biotechnology, and bioinformatics.

For each field of study we identified, we extracted the Wikipedia text of the page itself, as well as all of its citations. We then linked as many citations as possible to their titles and abstracts; these links could be established using the citations' DOI, Semantic Scholar ID, PubMed ID (PMID), or PubMed Central ID (PMC) and our dataset of scholarly literature. This gave us access to the cited publications' titles and abstracts, which we included in the ultimate text for each field. We also extracted each field of study mention in the text to use in our entity embeddings.

Using the extracted text, we then followed the algorithm described in Toney and Dunham (2022) to compute our document and entity embeddings for each field of study. With these embeddings, we were able to use cosine similarity to calculate a similarity score between each document in our corpus and each field of study. With 207,231,266 publications, and 1,110 fields, this gave us 230,026,705,260 initial scores.

However, while it is reasonable to have scores for all publications for all L0 and L1 fields, the same is not true for our L2 and L3 fields. This is because our L0 and L1 fields are comprehensive, and our L2 and L3 fields are not. If a publication receives its highest score for a particular L1 field, we can be reasonably confident it is related to that field, because our L1 fields are intended to broadly cover the scope of the scientific literature; the topic the publication discusses should be among our L1 fields and so its most similar embedding should be something actually relevant to it. But for our L2 and L3 fields, the field most similar to a publication may still not be similar at all. This is the challenge of building a non-comprehensive hierarchy; however, the alternative is to build a comprehensive hierarchy by hand – which is difficult and potentially unrealistic – or build a comprehensive hierarchy in an automated fashion – which leads to less intuitive results and is prone to error. Instead, we have chosen to create a method to eliminate unrelated results from our L2 and L3 scores.

After evaluation, our technique here is to rely on the L0 and L1 hierarchy. While one of the advantages of fields of study is their flexibility – publications can fall under multiple L0 and L1 fields – we ultimately believe most publications are unlikely to directly fall under more than a small number of disciplines. For that reason, we require any publication assigned an L2 or L3 field to have that L2 or L3 field's parent L0 field as one of their top two L0 fields, and its parent L1 field as one of their top three L1 fields. To provide an example, if a publication's highest-scoring L2 and L3 fields were "cryptography" and "differential privacy," we would expect that one of its two top-scoring L0 fields was "computer science" and one of its three top-scoring L1 fields was "computer security."

4 Results

Our final dataset included 1,110 fields, with 19 at L0, 280 at L1, 107 at L2, and 706 at L3. The smaller number of L2 fields as compared to L1 fields is explained by the narrowed scope at L2 – L2 fields don't cover the full scientific literature. The 207,231,266 publications over which fields were calculated were primarily in English, as the model was built based on English-language Wikipedia articles and their citations, but we also imputed scores for publications that had enough citation-based neighbors whose field scores we were able to calculate.

Our fields were generally based directly on individual, full Wikipedia articles and their linked citations. However, in certain cases where Wikipedia articles didn't align directly to the field in the taxonomy we wanted to cover, or the Wikipedia article included information that was likely to overlap multiple fields, we combined multiple articles or took specific sections of articles to develop our scores instead. In these cases we still used the article's citations, but limited ourselves to the citations of the portions of the articles we used. There were five fields that combined multiple articles and fifteen that used specific sections of articles.

When extracting references from articles, we focused on identifiable and linkable references in the scientific literature, ones with identifiers that we could connect to our dataset of scholarly literature. Of the 1,110 fields in our dataset, 949 of them had at least one such reference; for the others, we used just the Wikipedia text itself. The average length of the Wikipedia text for fields was 16,478 characters. The average length of the combined reference text, for fields with references, was 41,087 characters.

The distribution of our dataset among our level zero fields can be seen in Figure 1.

4.1 Field Representation Evaluations

As in Toney and Dunham (2022), we evaluate the resulting field representations by comparing their pairwise cosine similarities, with the expectation that vectors for closely-related fields should be proximal in the embedding space. Figure 2 shows the cosine similarity for each pair of level-zero fields of study. We expect, for example, that fields like computer science and engineering or business and economics should have relatively high cosine similarities, and they do; fields that are less related like biology and political science have relatively low cosine similarities.

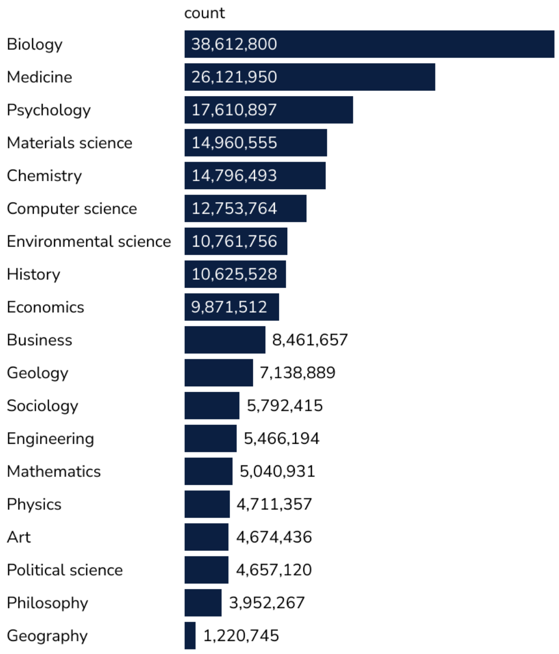

Figure 1: Counts of publications by their top level zero field.

In Figure 3, we inspect the relative position of fields in the embedding space using t-Distributed Stochastic Neighbor Embedding (t-SNE) to locate the 250-dimensional field embeddings in a 2-D plane. After dimensionality reduction, we can see that among subfields of computer science, the closest subfield to artificial intelligence is human-computer interaction. Meanwhile the related subfields of computer security, computer networks, and operating systems all appear near each other in the t-SNE plot. Similarly intuitive clusterings can be found in the t-SNE plot for the L2 and L3 subfields under artificial intelligence. For example, the nearest subfield to computer vision is gesture recognition, and we observe a clustering of neural networks, bio-inspired computing, and neuromorphic engineering.

4.2 Venue Matching

As one of our methods to evaluate our resulting fields, we produced field score outputs for a set of paper selected from conferences and journals that were focused on specific topics that were the same as or nearly identical to the fields themselves. So, for example, for our "human-robot interaction" field we looked at the *ACM/IEEE International Conference on Human Robot Interaction*, and for our "biometrics" field we examined publications

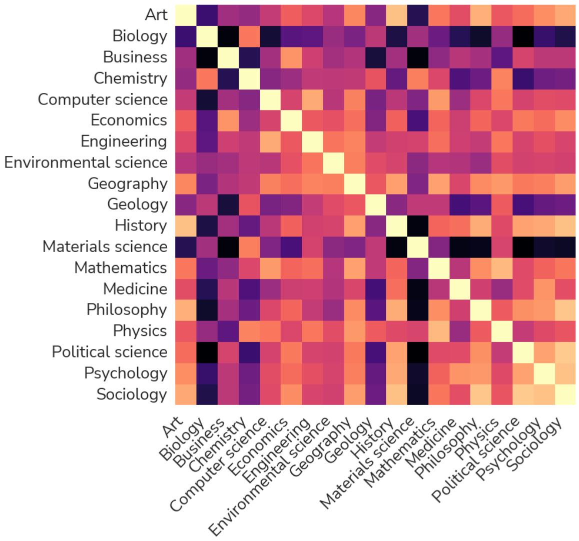

Figure 2: L0 Fields of Study cosine similarity heatmap.

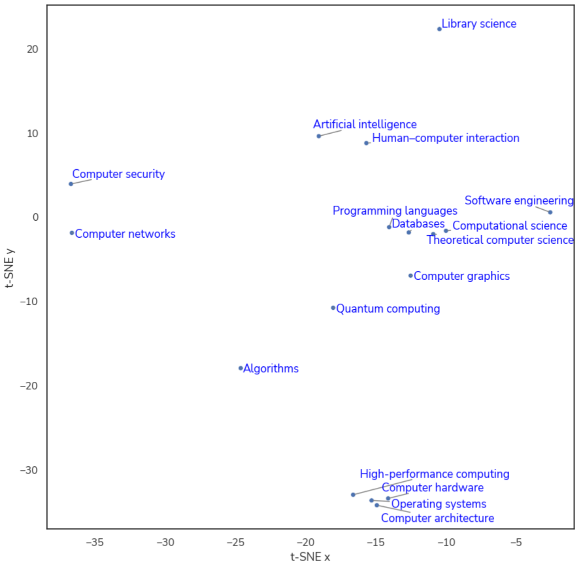

Figure 3: Computer science subfields t-SNE plot.

from the *International Joint Conference on Biometrics (IJCB)*.

This gave us a set of publications that we believed, with relatively high probability, should get high scores in specific fields of study. Directly matching conferences or journals did not exist for every field in our taxonomy, but we created an example subset, which enabled us to examine our results across a range of our new fields. Ultimately this subset included 55 conferences or journals covering 41 of our fields of study, at both level two and level three.

We then evaluated the fields of study scores on publications from those venues, looking to see how high our expected fields scored. We didn't anticipate that our expected field would always be the highest-scoring field; many of our fields have heavy overlap and many publications submitted to venues, even focused ones, touch on multiple areas. For example, one of the fields we selected to evaluate was "ethics of artificial intelligence," looking at the *Artificial Intelligence, Ethics, and Society (AIES)* conference. However, many publications there, not surprisingly, received their highest scores instead in "algorithmic bias," "fairness," "regulation of artificial intelligence," "explainable artificial intelligence," or even "AI safety." These are different but related topics, and ones that can all show up at the same venue, even if its top-level theme matches our field. Similarly, it was not uncommon for cross-topic publications to appear with their other topic (i.e. not the one from the venue they were in) as the top score, but to have the venue-relevant field be one of their other highest-scoring fields. One example would be a publication like "On Demographic Bias in Fingerprint Recognition" (Godbole et al., 2022), which appeared in a biometrics venue, but was marked as a paper about digital forensics. This overlap is actually one of the advantages of field scores, and of multi-label scoring systems in general, as it allows us to identify publications that naturally fall into multiple categories rather than just one.

Because of this natural tendency of publications to fall into multiple categories, we did not evaluate publications based solely on whether their highest-scoring field matched our expected venue, but instead considered whether one of their top five fields matched. With this evaluation, we found that 56.4% of our top five assigned fields matched their expected field based on their venue. We also identified highly-similar and overlapping fields (e.g. "algorithmic bias" and "fairness") and determined that 81.1% of our top five assigned fields matched either the expected field or one of its comparable fields.

In addition to looking at venues whose work aligned with our new fields, we also identified some venues whose work did not, selecting publications on art history, architecture, theology, sociology, chemistry, psychology, and statistics. We picked these fields to provide a variety that would give us areas with both greater and lower likelihood of plausible cross-disciplinary overlap with our new fields. We then manually labeled a sample of the publications that were assigned L2 and L3 fields within these disciplines to better understand if these assignments made sense or were in error, selecting ten publications from each field, or all publications in the field if there were fewer than ten assigned. This gave us a sample of 56 total publications, of which 38 (or 67.9%) were assessed as correctly assigned to our L2 and L3 fields. These cross-disciplinary publications are some of the most difficult to identify.

4.3 Comparisons to Other Work

The Field of Research Classification (FoRC) Shared Task at the Natural Scientific Language Processing Workshop (NSLP) 2024 (Ahmad et al., 2024b) provides another source of ground-truth labels for evaluation purposes. The shared task provides two datasets, one of which is a good analogue for our work: 1,500 papers from the ACL Anthology annotated using Taxonomy4CL, which defines 170 topics and subtopics of computational linguistics.

To evaluate our classifier against the labels for the shared task, we created a crosswalk from Taxonomy4CL to our own fields of study. In Taxonomy4CL, there are 44 top-level, 105 level-two, and 21 level-three topics. Among these, 33 have direct counterparts in our fields of study taxonomy, most of which have identical names. We subsetted the ACL Anthology papers from the FoRC shared task to those receiving any of our 33 intersecting labels, and then compared their top-scoring fields to their Taxonomy4CL labels.

In this evaluation, we found (micro) precision of 0.60 and recall of 0.60. For reference, the top-scoring submission for the shared task (Bashyam and Krestel, 2024) scored the evaluation set with (micro) precision of 0.44 and recall of 0.76. These metrics are not directly comparable to ours, after our restriction of the evaluation set to a subset of papers, but our purpose in evaluating against the Taxonomy4CL labels was only to assess the validity of our field labels, not to attempt the shared task. Relatively high performance against the ground truth from the shared task provides some evidence of our predictions' validity.

5 Conclusion

Extending our fields of study methodology to enable the creation of granular fields in subject-specific areas allows for much more detailed bibliometric analysis of publication data. Our methodology for doing so is repeatable, extensible, and relies on public resources like Wikipedia, citations from Wikipedia to publication data, and publicly available taxonomies from academic conferences and journals, as well as the expertise of the academic community. Our zero-shot approach requires no annotation or training data, making it extremely accessible, and uses fast, cheap embedding techniques and similarity metrics that can be run on a personal computer. Nonetheless it produces high-quality results across hundreds of fields.

One limitation of our current approach is our focus on English-language results. We have explored using Wikipedia pages in other languages to produce the same results, but more thorough evaluation is needed to properly assess the impact of using alternative pages, embedding models, and citation sets. In the future, we would like to extend to at least some of the most common languages in use in the publication literature. In the meantime, we have imputed scores for a subset of non-English publications that have direct citation links to English-language works.

It is possible that the most cutting-edge or niche fields may not appear in Wikipedia, either because they do not meet the notability guidelines or because no volunteer has yet written them up. In future work, it may be worth exploring whether bringing in external field definitions and citations from other locations, like journal subcategories, might produce additional fields to fill in gaps. Perhaps quality results from such an approach could even be contributed back to Wikipedia as new pages. Despite these limitations, our approach provides a valuable new technique for focused bibliometric analysis. The taxonomy, classifications, and code are available on GitHub.[1]

Acknowledgments

The authors would like to thank Autumn Toney-Wails for her valuable work on the previous iteration of this project, our anonymous reviewers for their insightful feedback, and colleagues at the Center for Security and Emerging Technology at Georgetown University for their feedback and support.

References

Raia Abu Ahmad, Ekaterina Borisova, and Georg Rehm. 2024a. Forc@ nslp2024: Overview and insights

[1] https://github.com/georgetown-cset/fields-of-study-pipeline

from the field of research classification shared task. In *Proceedings of the 1st International Workshop on Natural Scientific Language Processing and Research Knowledge Graphs (NSLP 2024)*.

Raia Abu Ahmad, Ekaterina Borisova, and Georg Rehm. 2024b. FoRC4CL: A fine-grained field of research classification and annotated dataset of NLP articles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7389–7394, Torino, Italia. ELRA and ICCL.

Lakshmi Rajendram Bashyam and R Krestel. 2024. Advancing automatic subject indexing: Combining weak supervision with extreme multi-label classification. In *Proceedings of the 1st International Workshop on Natural Scientific Language Processing and Research Knowledge Graphs (NSLP 2024). Hersonissos, Crete, Greece*, volume 27.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Akash Godbole, Steven A Grosz, Karthik Nandakumar, and Anil K Jain. 2022. On demographic bias in fingerprint recognition. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE.

Kelsey MacMillan and Sergey Feldman. 2023. Announcing s2fos, an open source academic field of study classifier.

OpenAlex. 2022. Automated concept tagging for openalex, an open index of scholarly articles.

Mark Sanderson and Bruce Croft. 1999. Deriving concept hierarchies from text. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 206–213.

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

Autumn Toney and James Dunham. 2022. Multi-label classification of scientific research documents across domains and languages. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 105–114, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. 2020. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies*, 1(1):396–413.