A Fine-grained citation graph for biomedical academic papers: the finding-citation graph

Yuan Liang
Queen Mary University
London, UK
yuan.liang@qmul.ac.uk

Roonak Rezvani
Exscientia
Oxford, UK
rrezvani@exscientia.co.uk

Massimo Poesio
Queen Mary University
London, UK
m.poesio@qmul.ac.uk

Abstract

Citations typically mention findings as well as papers. To model this richer notion of citation, we introduce a richer form of citation graph with nodes for both academic papers and their findings: the finding-citation graph (FCG). We also present a new pipeline to construct such a graph, which includes a finding identification module and a citation sentence extraction module. From each paper, it extracts rich basic information, abstract, and structured full text first. The abstract and vital sections, such as the results and discussion, are input into the finding identification module. This module identifies multiple findings from a paper, achieving an 80% accuracy in multiple findings evaluation. The full text is input into the citation sentence extraction module to identify inline citation sentences and citation markers, achieving 97.7% accuracy. Then, the graph is constructed using the outputs from the two modules mentioned above. We used the Europe PMC to build such a graph using the pipeline, resulting in a graph with 14.25 million nodes and 76 million edges.

1 Introduction

In recent years, the volume of biomedical literature has been constantly growing. More than 3000 articles are published every day on average and PubMed alone has a total of 29M articles as of January 2019 (Lee et al., 2019). This makes it difficult for experts to understand and assess the publications within a short amount of time.

Citations play a crucial role in academic papers, linking the new work to related research (Cohan et al., 2019). They can assist in evaluating research outputs (Yue and Wilson, 2004), and tracking the progression of research while predicting future directions (Prabhakaran et al., 2016). The citation network, a graph that records the citation relationship between papers, is commonly used in such studies (Gundolf and Filser, 2013; Hota et al., 2020; Zhao, 2020) .

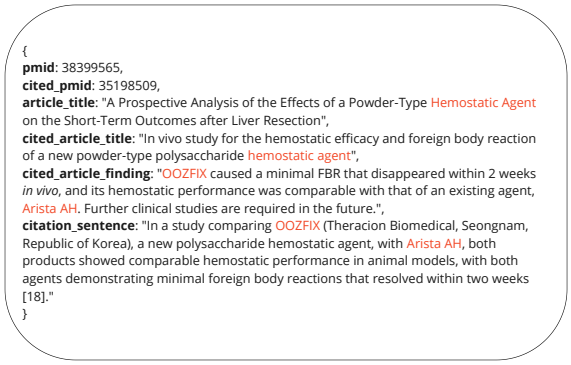

Figure 1: An example of the relation between paper and cited paper's finding through the citation sentence.

In recent years, many academic databases, which also can be regarded as academic citation networks, have been developed to facilitate detailed citation studies on biomedical publications. They provide basic information for hundreds of millions of academic papers and the citations between these documents. Some of these databases are commercial, like Clarivate's Web of Science (WoS) and Elsevier's Scopus. Others are open-source, in line with current trends. These include Microsoft Academic Graph (MAG) (Sinha et al., 2015), OpenCitations Index of CrossRef open DOI-DOI citations (COCI) (Heibi et al., 2019), Dimensions (Herzog et al., 2020), National Institutes of Health's Open Citation Collection (NIH-OCC) (Ian Hutchins et al., 2019), Semantic Scholar's Open Research Corpus (S2ORC) (Lo et al. 2020; Kinney et al. 2023). Some statistics about these databases are shown in Table 1.

From Table 1, it is clear that all databases, except S2ORC, lack inline citation contexts. These contexts provide information on what and why a paper cites information from other papers. For example, it may cite another paper's findings or refer to background or statistical information (Cohan et al., 2019). The findings of the paper are the most valuable output of the study. Only cita-

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 416–426

Database	Version	Publication	Citation	Access	Disciplines	Citation Contexts
Wos Core	2024	92M	2.2BN	Commercial	Multi	No
Scopus	2024	94M	2.4B	Commercial	Multi	No
MAG	2020-10	240M	-	Stop Serving	Multi	No
COCI	2023-01	77M	1.4B	Open Source	Multi	No
Dimensions	2024	140M	-	Application Needed	Multi	No
NIH-OCC	2024-04	37M	782M	Open Source	Health	No
S2ORC	2024	214M	2.49B	Application Needed	Multi	Yes

Table 1: A comparison between existing academic databases covering medical corpus.

tions of these findings can be used to evaluate the value of the publication and the research output. However, to understand whether a paper cites another paper's findings and which specific finding it refers to, the research findings need to be identified. A relationship between the citations and findings must also be established. An example can be seen in the Figure 1. Existing research on identifying research findings, such as the approach proposed by Wright et al. (2022) to extract sentences describing research findings and study their dissemination in scientific communication, can be helpful in this context. Motivated by the challenges of current databases and the existing research on finding identification, we propose the development of a fine-grained citation graph. This graph will involve both research findings and citation contexts. It will enable detailed evaluation and study of finding evolution from the citation perspective.

In this paper, we define the finding-citation graph first in section 3. Then, we outline the process of constructing the finding-citation graph using the European PMC dataset in section 4. We also evaluate the construction to ensure quality. The summary statistics of the graph and some interesting observations are presented in section 5.

2 Related Work

Constructing a fine-grained citation graph directly relates to cite-worthiness detection, and finding identification. We will briefly introduce these aspects in the following sub-sections. Since the biomedical large language model (LLM) has recently gained popularity and may be used in our project, it will also be introduced in the subsequent sub-sections.

Cite-Worthiness Detection Cite-worthiness detection involves identifying citation sentences in an academic paper. These sentences contain references to external sources cited within the paper. There are many different forms of citation, but the most common are:

- The topic is studied in previous work (Author1 et al. ###).
- The topic is studied in previous work [##].
- The topic is studied in previous work (##).
- The topic is studied using XXX (Author 1 et al. ###) and XXX (Author1 et al. ###) XXX.
- Author 1 et al, ### (year) performs XXX.

Sugiyama, Kumar, Kan, and Tripathi (2010) suggested the application of Support Vector Machines (SVMs) with diverse features for cite-worthiness detection. These features range from unigrams, bigrams, and the existence of proper nouns, to section information, classification of neighboring sentences, and orthographic checking. They designed a dataset using the ACL Anthology Reference Corpus (ACL-ARC) (Bird et al., 2008), applying regular expression patterns. Similarly, Färber et al. (2018b) carried out the same task using convolutional recurrent neural networks on an expanded dataset. The dataset incorporated three subsets: ACL-ARC (Bird et al., 2008), arXiv CS (Färber et al., 2018a), and Scholarly Dataset 2.

However, the datasets from these studies are confined to one or a limited number of domains and exhibit a high class imbalance. As per Färber et al. (2018b), only a tenth of all sentences hold at least one citation marker, leaving the remaining 90% without any. Furthermore, these studies lack an in-depth discussion on dataset creation and qualitative analysis.

In response to these issues, Wright and Augenstein (2021) introduced a dataset for spotting

citation-worthy sources across six domains. They detailed the process for creating the dataset and provided a qualitative analysis. However, their approach to dataset creation was limited to using regular expressions to identify the first and second citation forms mentioned above. Besides, the authors trained a set of baseline models on their dataset to evaluate performance and understand the complexity of the problem. The results of these models are displayed in Fig 2.

Method	P	R	F1
Logistic Regression	$46.65_{0.00}$	$64.88_{0.00}$	$54.28_{0.00}$
Färber et al. (2018b)	$49.57_{0.96}$	$65.56_{2.61}$	$56.41_{0.34}$
Transformer	$47.92_{0.78}$	$71.59_{1.74}$	$57.39_{0.10}$
BERT	$55.04_{0.66}$	$69.02_{1.33}$	$61.23_{0.21}$
SciBERT-no-weight	$\mathbf{65.94}_{0.37}$	$51.62_{0.53}$	$57.91_{0.30}$
SciBERT	$57.03_{0.50}$	$68.08_{1.03}$	$62.06_{0.15}$
SciBERT + PU	$49.46_{0.83}$	$\mathbf{82.12}_{1.40}$	$61.73_{0.27}$
Longformer-Solo	$57.21_{0.25}$	$68.00_{0.41}$	$62.14_{0.02}$
Longformer-Ctx	$59.92_{0.28}$	$77.15_{0.49}$	$\mathbf{67.45}_{0.06}$

Figure 2: Performance of models on the CITEWORTH dataset (Wright and Augenstein, 2021)

Finding Identification The process of pinpointing and extracting results or conclusions from an academic paper is known as finding identification. Prabhakaran, Hamilton, McFarland, and Jurafsky (2016) designed a Conditional Random Field (CRF) model that manages sentence-level sequence labeling, designating each sentence in the abstract a rhetorical role, including result and conclusion. Dernoncourt and Lee (2017) introduced a considerable sentence classification dataset, PubMed 200K RCT. This dataset, consisting of roughly 200,000 abstracts of randomized controlled trials (RCTs) and a total of 2.3 million sentences, labels each sentence with its rhetorical role, which includes the result and conclusion. Though it is limited to the RCT field, this dataset can help find identification. Inspired by the PubMed 200K RCT dataset, Wright et al. (2022) curated a dataset of 200K self-labeled abstracts from PubMed, with no field restrictions. Then, they fine-tuned a RoBERTa model (Liu et al., 2019) on this dataset, classifying each sentence in the abstract into categories such as result, conclusion, method, background, and others. The model achieved an F1 score of 92%, and when applied to the full text of papers, it performed well. Previous studies have overlooked the importance of certain sections of papers, such as the results and conclusion sections. These sections often contain important findings. Past studies mainly focus on finding extraction, neglecting the potential for finding generation. However, the development of large generative language models, like Llama (Touvron et al., 2023), now provides the opportunity for effective finding generation.

In the finding identification task, there is a subtask known as claim identification or argumentation mining exists. According to the definition from Achakulvisut et al. (2020b), a claim is (1) a statement declaring something as superior, (2) a statement proposing something new, or (3) a statement describing a new discovery or a new cause-effect relationship. The definition of a claim differs from that of a finding, being stricter and more precise. Nonetheless, some ideas from this research could be useful. Achakulvisut et al. (2020b) developed a tool for annotating claims and collected 1500 labeled abstracts (SciCE) from PubMed articles published from 2008 to 2018. These abstracts incorporate 11,702 sentences in total, with each sentence labeled as a claim or non-claim. This tool effectively tackles the issue of data scarcity in the task. They also constructed a new model incorporating transfer learning, which improved the F1 score by 14 percentage points compared to the baseline model without transfer learning. In 2023, Wei et al. undertook the same task, achieving a new state-of-the-art (SOTA) performance on the dataset using supervised contrastive learning and transfer learning, with an 87.45% F1 score. As observed, all these models operate on the abstract rather than the full-text article, and the shift to the full-text article still poses a challenge due to the writing structure of the complete publication.

Biomedical LLMs Back in 2018, ELMo (Peters et al., 2018) pioneered the use of a context-sensitive language model pre-trained on a huge data corpus. This sparked a wave of LLMs such as GPT (Radford and Narasimhan, 2018), BERT (Devlin et al., 2019), ERNIE (Zhang et al., 2019), GPT-2 (Radford et al., 2019), GPT3 (Brown et al., 2020), among others. These LLMs are incredibly useful for a variety of natural language processing (NLP) tasks. However, general-purpose LLMs, which are trained on resources like English Wikipedia and BookCorpus, often struggle with biomedical NLP tasks due to the numerous domain-specific terms and proper nouns. To counter this, many LLMs have been pre-trained on biomedical corpus

like PubMed abstracts and PubMed Central full-text articles, enhancing their performance in the biomedical field.

There are two main approaches to domain-specific neural language model paradigms: mix-domain pre-training and domain-specific pre-training from scratch. Mix-domain pre-training, such as BioBERT (Lee et al., 2019) and BlueBERT (Peng et al., 2019), begins with parameters from a general-purpose language model and adopts its vocabulary. On the other hand, domain-specific pre-training from scratch, like PubMedBERT (Gu et al., 2021), BioLinkBERT (Yasunaga et al., 2022), BioMedLM (Bolton et al., 2024), and Bioformer-8L (Fang et al., 2023), generates vocabulary and conducts pre-training using only the in-domain corpus. Models like PubMedBERT and BioMedLM have shown that domain-specific pre-training from scratch can outperform mix-domain pre-training.

3 The Finding-Citation Graph: Definition

Building on the work of Wright et al. (2022), we define a finding as a statement that describes a specific research outcome from a scientific study. We also describe a citation sentence as a sentence that references knowledge from other papers.

Subsequently, we define a finding-citation graph (FCG) as $G = (P, F, C, B)$, where P, F, C, and B represent sets of papers, findings, citations, and basic information respectively. A paper in this graph is an academic paper. A finding in this graph is a statement same as the above definition. A citation within the graph refers to instances where the citation sentence includes the findings of the cited paper, which we will now refer to as a useful citation. The basic information includes the author, journal, publication year, etc. of the paper containing the finding. The defined finding-citation graph is a heterogeneous graph and can be perceived as a variant of the citation graph, where the node is the paper and the relation is the citation.

4 Constructing the FCG

We now introduce our pipeline to construct the finding-citation graph (Figure 3), which allows us to analyze findings from the perspective of the citation network. The pipeline takes a Europe PubMed article in XML format as input and produces three types of information for each paper: basic information, all citation sentences, and all findings. This information is then used to construct the graph. The pipeline comprises four main modules, as follows:

- An XML parser is utilized to extract essential paper information and article content from the XML. The primary components of the article include the abstract and the full-text article, composed of various sections.
- The finding identification model aims to identify sentences that describe findings from the abstract, conclusion, and result sections.
- The citation sentence extraction module identifies sentences within the full-text article that contain citations and links the citation sentences with its cited paper.
- The final module is to build the finding-citation graph construction based on the output of the above three modules

The above four modules will be introduced in the following sections excluding the XML parser. Our parser was primarily based on an open-source PubMed parser (Achakulvisut et al., 2020a), with minor changes made to increase speed.

4.1 Finding Identification

This module includes two steps:

- Identify the sentences that discuss the findings, which are called finding sentences later.
- Generate findings based on the identified finding sentences, which are called findings later.

We performed the first step similar to Prabhakaran et al. (2016) and Wright et al. (2022), where the task was treated as a sentence classification task. They classified sentences in the abstract into five classes: background, result, conclusion, method, and objective. Sentences labeled as the result or conclusion can be considered findings sentences. To build the sentence classification model, we curated a dataset from self-annotated PubMed abstracts, as shown in Figure 4. After filtering for PubMed abstracts that met the set format, we obtained 206K suitable abstracts, comprising roughly 2.5 million labeled sentences. We then fine-tuned a RoBERTa model (Liu et al., 2019) on this curated dataset, achieving an accuracy score of 91% on a held-out 13.5% sample. This classifier was applied to the abstract sentences and other sections of the paper, like the results and conclusion, generating multiple finding sentences for each paper.

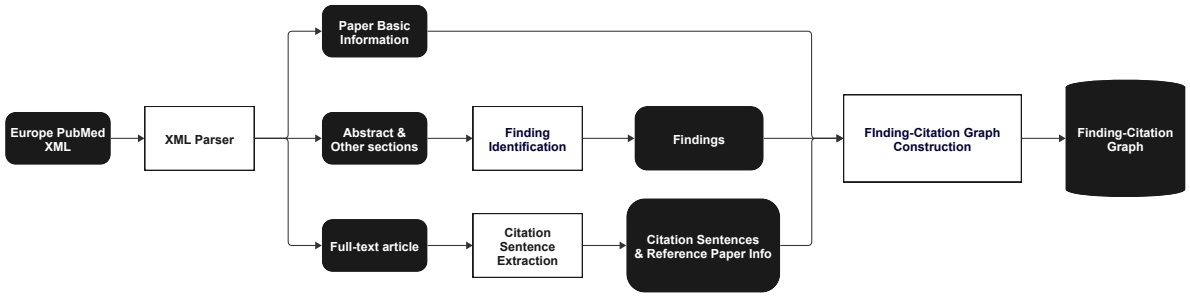

Figure 3: Finding-citation graph construction pipeline

We also experimented with LLMs to assess their potential as a substitute for the current fine-tuned model. Specifically, we utilized Gemma (Gemma Team et al., 2024) to assist us in categorizing the information in the abstract into the five classes mentioned above. To evaluate Gemma's performance, we compared the organized information for each class with self-annotated information, calculating similarity scores. When we set 0.5 as the similarity threshold, the accuracy was approximately 91%. As the performance is nearly the same, taking the time-consuming and resource-consuming into consideration, we chose to use fine-tuned RoBERTa model in our pipeline.

Rationale: Neonatal ibotenic acid lesion of the ventral hippocampus was proposed as a relevant animal model of schizophrenia reflecting positive as well as negative symptoms of this disease. Before and after reaching maturity, specific alterations in the animals' social behaviour were found.

Objective: In this study, social behaviour of ventral hippocampal lesioned rats was analysed. For comparison, rats lesioned either in the ventral hippocampus or the dorsal hippocampus at the age of 8 weeks were tested.

Methods: Rats on day 7 of age were lesioned with ibotenic acid in the ventral hippocampus and social behaviour was tested at the age of 13 weeks. For comparison, adult 8-week-old rats were lesioned either in the ventral or the dorsal hippocampus. Their social behaviour was tested at the age of 18 weeks.

Results: It was found that neonatal lesion resulted in significantly decreased time spent in social interaction and an enhanced level of aggressive behaviour. This shift is not due to anxiety because we could not find differences between control rats and lesioned rats in the elevated plus-maze. Lesion in the ventral and dorsal hippocampus, respectively, in 8-week-old rats did not affect social behaviour.

Conclusions: The results of our study indicate that ibotenic acid-induced hippocampal damage per se is not related to the shift in social behaviour. We favour the hypothesis that these changes are due to lesion-induced impairments in neurodevelopmental processes at an early stage of ontogenesis.

Figure 4: An example of a self-annotated PubMed abstract from PubMed PMID:10435405.

It is important to note that these finding sentences may contain overlap information with each other and may discuss multiple discoveries from the paper. Additionally, not all findings carry equal importance to the article. Our next step involved generating multiple findings for each paper, keeping these points in mind. We used a combination of scientific sentence-BERT (Wright et al., 2022) and the Affinity Propagation clustering method to eliminate duplicate sentences and select central sentences as representative findings. This approach yielded multiple findings (maximum 3) from each paper. Afterward, we computed the similarity score between each finding and the title of its corresponding article. This score is considered as the importance score for each finding within its respective paper. Consequently, we obtained multiple findings for each paper along with an importance ranking score. This procedure is illustrated in Figure 5. In order to know how the multi-finding module performs, we randomly sampled 100 articles with 241 findings and got 80% accuracy. We found that some errors originated from the abstract's conclusion sentence, which did not accurately represent the actual conclusions and simply offered a concluding sentence without any useful information.

4.2 Citation Sentence Extraction

The task involves identifying sentences in the article that reference external knowledge from other papers. Unlike other researchers such as Sugiyama et al. (2010) and Färber et al. (2018b), who employed binary classification models for this task, we used a simpler yet effective method: the regular expression. We addressed three formats of citation using this method, shown below. The use of regular expression simplified the process of linking the citation sentence with its cited paper. This was based on the citation marker and reference information derived from the XML. Consequently, we obtained citation sentences along with the information of the cited paper for each article.

- The topic is studied ... (Author1 et al. ###).
- The topic is studied ... [###].
- The topic is studied ... (###).

To evaluate how the module performed and maximize the use of the open-source dataset, we designed the following two-step evaluation method.

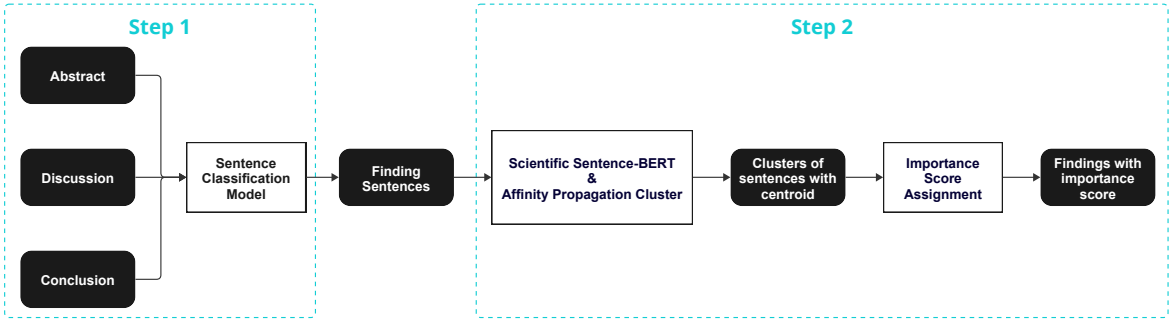

Figure 5: Finding Identification Procedure

The first is a paper-level evaluation, which checks the accuracy of the citation relations, i.e., the PMID-PMID relation. Same as Liang et al. (2021), we utilized the PMC dataset in PubMed Baseline as the gold standard for this evaluation. Both the original and filtered PMID-PMID relations of the module were evaluated, with the latter excluding references not found in the Europe PMC dataset. In order to do the comparison with other open-source datasets, we used the same evaluation metrics as Liang et al. (2021), which are precision, recall, F1-score, and accuracy. It should be noted that only articles covered by the data source were included in the evaluation process. The evaluation results can be seen in Table 2. Even though our performance is not bad, our precision and recall are not the best among all the databases because the citation relations were based on citation sentences and markers, not the reference list. This discrepancy may lead to errors and losses in citation relations.

The second evaluation is to assess the correctness of the tuple, $(citing\_pmid, citation\_sentence, cited\_pmid)$. This formed the final output of the module. There is no other database containing the citation sentences on the PubMed corpus, except for S2ORC (Lo et al., 2020). However, the performance evaluation from Step 1 indicates that the S2ORC database did not perform well. Moreover, the S2ORC paper (Lo et al., 2020) does not provide a significant evaluation of the citation sentence, so we do not use it for evaluation. We randomly sampled 350 tuples and achieved 97.7% accuracy. We conducted an analysis to determine why certain tuples are incorrect. We found that some errors arise from mismatches between the description citation marker and the basic information of the cited paper. Other errors occur when the citation sentences are correctly identified, but the PMID of the cited paper is lost. This largely confirms our previous analysis above that the citations are based on citation sentences and citation markers can lead to errors and losses in this module.

4.3 Finding-Citation Graph Construction

So far, we have collected multiple findings for each paper, along with their importance scores and citation sentences with basic information about the cited paper. Using this information, we can create the finding-citation graph as outlined in Section 3.1. As we construct the graph based on a closed dataset, Europe PMC, the references without PMID or not in the closed dataset were dropped.

The graph comprises two types of nodes: articles and findings. Each article node has some basic attributes such as authors, journal, paper PMID, title, and publication year. In contrast, the finding node has no other attributes. The graph also contains two kinds of edges. The first represents the relationship between an article and its findings, with the importance score as an attribute. The second represents the citation relationship between an article and the findings produced by the cited paper, with the citation sentence and similarity score as attributes. We calculate the similarity score using a fine-tuned scientific sentence-BERT (Wright et al., 2022). This approach helps us determine whether a citation sentence contains the findings of another paper and assess the usefulness of each citation. A simplified view of the graph can be seen in Figure 6.

5 Experiments

We utilized Europe PMC articles in XML format for our experiment. Europe PMC is an open-source, global biomedical literature repository that houses life science articles, preprints, micropublications, books, patents, and clinical guidelines from around the world. Up to Feb 2024, it holds over 40 million

Metrics	COCI.Updated	Dimensions	NIH-OCC	S2ORC	S2ORC_new	Our_O	Our_D
Precision	98.82%	99.60%	99.9%	97.66%	77%	94.32%	92.35%
Recall	85.18%	98.80%	98.99%	79.00%	25.4%	89.65%	93.05%
F1-score	90.95%	99.07%	99.34%	86.27%	34.5%	89.32%	90.7%
Accuracy	15.60%	81.55%	89.08%	5.86%	1.03%	34.4%	55.37%

Table 2: The evaluation of COCI.Updated, Dimensions, NIH-OCC, and S2ORC are from Liang et al. (2021). Our main comparison is S2ORC, which is the most similar database to use and includes citation contexts, so we evaluated S2ORC on the latest version again. It is not only for the comparison but also for the confirmation of our comparison. The Our_O and Our_D are the original and filtered PMID-PMID relations respectively. When we did the evaluation on the filtered PMID-PMID relations, we did the same filter to the gold dataset.

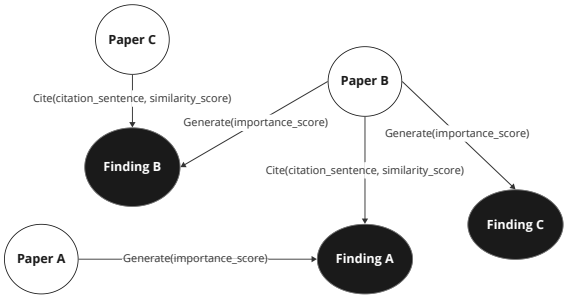

Figure 6: A simplified view of the finding-citation graph

Type of Article	Count(%)
Europe PMC XML	6M (100%)
Successfully parsed	5.75M (95.8%)
With PMCID	5.75M (95.8%)
With PMID	5.75M (95.8%)
With Abstract	4.83M (80%)
With Paragraph	5.75M (95.8%)
With References	5.21M (86.8%)

Table 3: Statistics of the XML Parser output.

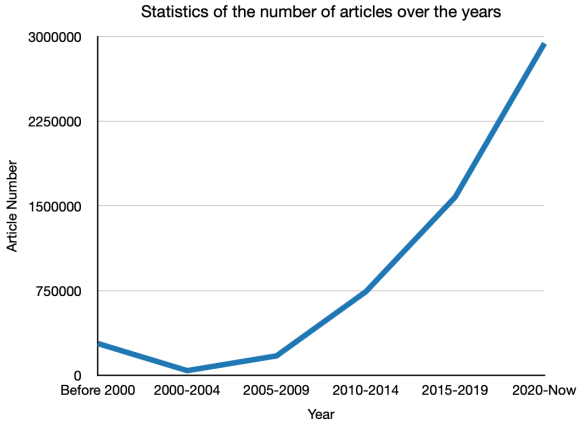

Figure 7: Statistic of the number of articles over the years in Europe PMC

abstracts and over 9.6 million full-text articles. Of these, nearly 6 million full-text open-access articles are available in XML format via the Europe PMC web services or FTP site.

The XML Parser in our pipeline is used to parse all the open-access articles mentioned above. The statistical results of the output of this module are presented in Table 3.

Next, the parsed text is processed by the finding identification module and the citation sentence extraction module. In terms of finding identification results, approximately 4.25 million articles have at least one finding. On average, we obtained 2.4 findings for each article that had findings, totaling around 10 million findings. For citation sentence extraction results, roughly 3.69 million articles have at least one citation sentence. We obtained 56 citation sentences on average for each article with citation sentences, amounting to approximately 200 million citation sentences in total. After dropping the citations not in Europe PMC, roughly 3.28 million articles have at least one citation sentence, with 20 citation sentences on average for each article and 67 million citation sentences in total.

Finally, the similarity scores were calculated based on these findings and citation sentences. These findings, citation sentences, and similarity scores are then utilized to create the finding-citation graph. We obtained 14.25 million nodes in total, consisting of 4.25 million article nodes and 10 million finding nodes. We got 77 million edges in total, of which there are 67 million edges representing citation relationships.

6 Discussion

6.1 Findings not in the abstract

From the literature review, it is clear that most previous studies primarily focus on identifying finding sentences from abstracts, often neglecting other sections. In our proposal to identify multiple findings, we are interested in determining how many find-

ings are not included in the abstract, meaning the sentences containing these findings are not found in the abstract. From the 10 million findings we identified, we discovered that nearly 44% of the findings are not mentioned in the abstract. This percentage is slightly larger than expected. However, it aligns with our understanding that the abstract typically only describes the main findings, while other sections may discuss additional or side findings.

6.2 Distribution of similarity score between findings and citation sentences

The sentence embedding similarity score is utilized to measure how much information from the cited paper's findings is contained within the citation sentence. The larger it is, the more information it contains. The smaller it is, the less information it contains. Understanding the distribution of such similarity scores can help us determine how the findings are cited. The distribution can be seen in Figure 8. From the figure, it is clear that nearly half of the similarity scores are lower than 0.4. This suggests that half of the citation sentences contain less information about the findings of the cited papers. It meets our experience that most of the citations are used in the literature review section and for providing background information.

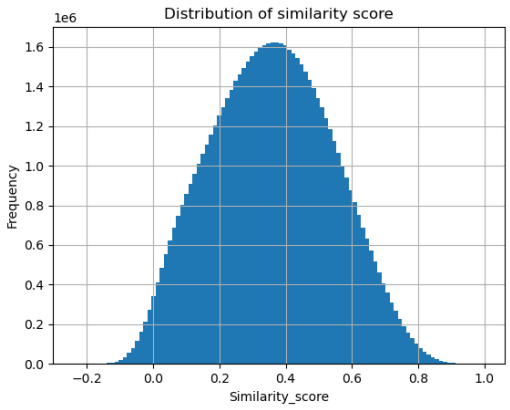

Figure 8: Distribution of the similarity score between citation sentences and cited findings

6.3 Limitations

Currently, our method only extracts sentences with the sentence marker for the citation, without considering the citation span. This approach might lead to some errors in matching the citation sentence with the findings.

Besides, the citation relations are based on the citation sentence and citation marker extraction, which would lead to error and loss to the graph.

Moreover, the citation graph is built using a closed dataset, specifically the Europe PubMed dataset. This approach excludes citations and articles not found in this dataset, inevitably leading to an incomplete graph.

The mentioned limitations above will help us identify areas where we can improve our graph optimization strategies in the future.

7 Conclusion

We introduce a new fine-grained citation graph, the finding-citation graph. Unlike the traditional citation graph which only contains papers as nodes, the finding-citation graph also includes the findings, representing the results of the academic papers. This graph facilitates more detailed studies at the finding level, such as evaluating findings and tracking the progression of research.

We also present a new pipeline for constructing this graph. This pipeline mainly consists of three modules: finding identification, citation sentence extraction, and graph construction. As there is no such pipeline to build the finding-citation graph, we design an evaluation to confirm the graph's quality. The finding identification module achieved 91% accuracy for finding sentence identification and 80% accuracy for multiple findings. The citation sentence extraction module got a 90% F1-score on the paper-level evaluation and 97.7% accuracy on the tuple-level evaluation. The outputs of the two modules are used to construct the graph and confirm its quality.

Finally, we built a finding-citation graph using Europe PMC. Our graph comprises 14.25 million nodes, with 4.25 million being academic papers and the rest being findings from those papers. It also includes 76 million edges, with 66 million representing citation relations.

The definition and creation of the FCG is an essential step for our future research. We plan to use it to assess research findings from a citation perspective and pinpoint future research directions at the finding level.

Ethics Statement

The paper considers the introduction of a new citation network, a finding-citation network, and a pipeline to construct such a graph. We did not work

with limited datasets and only used open-source datasets.

Acknowledgements

This work was supported by the UKRI Biotechnology and Biology Sciences Research Council [BB/X511833/1], Digital Environment and Research Institute (DERI), the Queen Mary University of London, and Exscientia.

We thank Arkaitz Zubiaga and Daniel Crowther for their valuable feedback and suggestions on the project. We also thank the Semantic Scholar team for assisting with data access.

References

Titipat Achakulvisut, Daniel Acuna, and Konrad Kording. 2020a. Pubmed parser: A python parser for pubmed open-access xml subset and medline xml dataset xml dataset. *Journal of Open Source Software*, 5(46):1979.

Titipat Achakulvisut, Titipat Bhagavatula, Daniel Acuna, and Konrad Kording. 2020b. Claim extraction in biomedical publications using deep discourse model and transfer learning.

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

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D. Manning. 2024. Biomedlm: A 2.7b parameter language model trained on biomedical text.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3586–3596, Minneapolis, Minnesota. Association for Computational Linguistics.

Franck Dernoncourt and Ji Young Lee. 2017. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 308–313, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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

Li Fang, Qingyu Chen, Chih-Hsuan Wei, Zhiyong Lu, and Kai Wang. 2023. Bioformer: an efficient transformer language model for biomedical text mining.

Michael Färber, Alexander Thiemann, and Adam Jatowt. 2018a. A high-quality gold standard for citation-based tasks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Michael Färber, Alexander Thiemann, and Adam Jatowt. 2018b. *To Cite, or Not to Cite? Detecting Citation Contexts in Text*, pages 598–603.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Dixon, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas,

Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu-hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Katherine Gundolf and Matthias Filser. 2013. Management research and religion: A citation analysis. *Journal of Business Ethics*, 112:177–185.

Ivan Heibi, Silvio Peroni, and David Shotton. 2019. Software review: Coci, the opencitations index of crossref open doi-to-doi citations. *Scientometrics*, 121(2):1213–1228.

Christian Herzog, Daniel Hook, and Stacy Konkiel. 2020. Dimensions: Bringing down barriers between scientometricians and data. *Quantitative Science Studies*, 1:387–395.

Pradeep Kumar Hota, Balaji Subramanian, and Gopalakrishnan Narayanamurthy. 2020. Mapping the intellectual structure of social entrepreneurship research: A citation/co-citation analysis. *Journal of Business Ethics*, pages 1–26.

B. Ian Hutchins, Kirk L. Baker, Matthew T. Davis, Mario A. Diwersy, Ehsanul Haque, Robert M. Harriman, Travis A. Hoppe, Stephen A. Leicht, Payam Meyer, and George M. Santangelo. 2019. The nih open citation collection: A public access, broad coverage resource. *PLOS Biology*, 17(10):e3000385.

Rodney Michael Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F.Q. Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler Murray, Christopher Newell, Smita R Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, A. Tanaka, Alex D Wade, Linda M. Wagner, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine van Zuylen, and Daniel S. Weld. 2023. The semantic scholar open data platform. *ArXiv*, abs/2301.10140.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Zhentao Liang, Jin Mao, Kun Lu, and Gang Li. 2021. Finding citations for pubmed: a large-scale comparison between five freely available bibliographic data sources. *Scientometrics*, 126(12):9519–9542.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

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

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

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

Vinodkumar Prabhakaran, William L. Hamilton, Dan McFarland, and Dan Jurafsky. 2016. Predicting the rise and fall of scientific topics from trends in their rhetorical framing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1170–1180, Berlin, Germany. Association for Computational Linguistics.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

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

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, page 243–246, New York, NY, USA. Association for Computing Machinery.

Kazunari Sugiyama, Tarun Kumar, Min-Yen Kan, and Ramesh C. Tripathi. 2010. Identifying citing sentences in research papers using supervised learning. In *2010 International Conference on Information Retrieval Knowledge Management (CAMP)*, pages 67–72.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Xin Wei, Md Reshad UI Hoque, Jian Wu, and Jiang Li. 2023. Claimdistiller: Scientific claim extraction with supervised contrastive learning. 3451:65–77.

Dustin Wright and Isabelle Augenstein. 2021. CiteWorth: Cite-worthiness detection for improved scientific document understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1796–1807, Online. Association for Computational Linguistics.

Dustin Wright, Jiaxin Pei, David Jurgens, and Isabelle Augenstein. 2022. Modeling information change in science communication with semantically matched paraphrases. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1783–1807, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.

Weiping Yue and Concepción S. Wilson. 2004. Measuring the citation impact of research journals in clinical neurology: A structural equation modelling analysis. *Scientometrics*, 60(3):317–332.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Qihang Zhao. 2020. Utilizing citation network structure to predict citation counts: A deep learning approach.