

Data Mining and Information Extraction Methods for Large-Scale High-Quality Representations of Scientific Publications

Zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften
(Dr.-Ing.)

von der KIT-Fakultät für Wirtschaftswissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte
Dissertation

von

M. Sc. Tarek Saier

Tag der mündlichen Prüfung:

22. April 2024

Referent:

Prof. Dr. Michael Färber

Korreferent:

Prof. Dr. Adam Jatowt



This document is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0): <https://creativecommons.org/licenses/by/4.0/deed.en>

Abstract

This dissertation addresses the challenge of generating high-quality, machine-readable representations of scientific publications at a large scale. Structured data representing scientific publications is the basis for vital infrastructure in academia, such as academic search and bibliometric performance indicators. Generating such data involves information extraction from publications' natural language content, which makes it a challenging and error-prone process. Existing extraction methods and the data they produce are limited in several ways. This is problematic, because it means that applications and research based on currently available data are of limited scope and validity.

Among the limitations of currently available methods and data, three areas are of particular importance due to their relevance in the academic context. (1) *Citation networks* are a key characteristic of scientific literature, and are vital for common use cases such as trend analyses and recommender systems. Despite this importance, citation networks of widely used data sets are highly incomplete. (2) *Language coverage*: science is a global and therefore inherently multi-lingual endeavor. Despite a growing awareness of this, important platforms, approaches, and data sets in the scholarly domain are still limited to English publications only. (3) *Research artifacts*, such as methods and data sets, become more and more important, as science is increasingly driven by curated data and algorithmic processing. Fine-grained representations of research artifacts bear large potential for applications like faceted academic search and automated reproduction. However, existing extraction methods only yield shallow representations of research artifacts, not sufficient for these use cases.

To address these issues, we develop data mining and information extraction approaches, that enable the creation of machine-readable publication corpora. We furthermore quantify the improvements we achieve in terms of data quality in each area of limitation. In particular, we make the following contributions. As the foundation of our research, we develop a method for creating a large-scale corpus of interlinked, full-text documents from publications' \LaTeX sources. Applying our method to all of arXiv.org, we create the first corpus of interlinked publications with extensive coverage in physics, mathematics, and computer science. Utilizing our corpus, we further present approaches yielding advances in all of the three aforementioned areas of limitation. (1) We develop a methodology for linking bibliographic references, which achieves state-of-the-art citation network completeness. Based on this, we perform novel types of citation analyses. (2) We present a method for identifying cross-lingual citations and, utilizing it, perform the largest analysis of this type of citation to date. Through our analysis, we are able to identify challenges for integrating non-English publications. (3) We develop information extraction approaches for fine-granular representations of research artifacts and their parameters. Our methods achieve an improvement over strong baselines, and their utilization enables novel types of analyses and applications.

Overall, our approaches address key shortcomings of existing methods for the creation of structured data representing publications. Through their use, we achieve significant improvements in terms of data quality. For each of our approaches, we demonstrate its viability and benefits through evaluations and practical large-scale applications. Our methods have already been adopted in several parts of the research community, which further confirms their utility.

Zusammenfassung

Diese Dissertation befasst sich mit Methoden zur automatisierten Extraktion qualitativ hochwertiger, strukturierter Daten aus wissenschaftlichen Publikationen. Strukturierte Daten über wissenschaftliche Publikationen ermöglichen essenzielle Elemente des akademischen Alltags, wie beispielsweise akademische Suchdienste und bibliometrische Kennzahlen. Die Erzeugung derartiger Daten erfordert die Extraktion von Informationen aus den natürlichsprachlichen Inhalten von Publikationen—ein Prozess, der komplex und fehleranfällig ist. Bestehende Extraktionsmethoden, und somit auch die damit gewonnenen Daten selbst, weisen diverse Mängel auf. Das ist problematisch, da folglich Anwendungen und Forschung basierend auf den derzeit verfügbaren Daten in ihrer Anwendbarkeit und Validität beschränkt sind.

Unter den Mängeln derzeit verfügbarer Methoden und Daten sind drei Bereiche von besonderer Bedeutung im akademischen Kontext. (1) *Zitiernetzwerke* sind ein essenzieller Teil wissenschaftlicher Literatur und spielen eine zentrale Rolle bei Trendanalysen und Empfehlungssystemen. Trotz ihrer Wichtigkeit sind Zitiernetzwerke viel verwendeter Datensätze hochgradig unvollständig. (2) *Sprachabdeckung*: Wissenschaft ist ein globales und damit inhärent multilinguales Unterfangen. Trotz zunehmender Anerkennung dieses Umstandes sind wichtige akademische Plattformen, Methoden und Datensätze auf englischsprachige Publikationen beschränkt. (3) *Forschungsartefakte*, wie beispielsweise Methoden und Datensätze, werden zunehmend wichtig, da Forschung mehr und mehr durch Daten und deren algorithmische Verarbeitung vorangetrieben wird. Feingranulare Daten über Forschungsartefakte können vielversprechende Anwendungen wie Facettensuche und automatisierte Replikation ermöglichen. Bestehende Extraktionsmethoden erfassen allerdings nur grobe Daten über Forschungsartefakte, die für derartige Anwendungen nicht ausreichen.

Wir adressieren diese Mängel durch die Entwicklung von Data-Mining- und Informationsextraktions-Methoden, welche die Erstellung maschinenlesbarer Publikationskorpora ermöglichen. Zusätzlich quantifizieren wir die damit erzielten Verbesserungen in der Datenqualität. Die hierfür umgesetzten Forschungsbeiträge sind wie folgt. Als Basis unserer Forschung entwickeln wir eine Methode zur Erstellung eines großen Korpus von verknüpften Volltextdokumenten auf Basis von \LaTeX -Quelldateien. Durch Anwendung unserer Methode auf der Gesamtheit von arXiv.org, erstellen wir den ersten Korpus verknüpfter Publikationen mit umfangreicher Abdeckung in der Physik, Mathematik und Informatik. Aufbauend auf diesem Korpus entwickeln wir Ansätze, die Fortschritte in allen der drei zuvor erwähnten Mängelbereichen erzielen. (1) Wir entwickeln Methoden für die Verknüpfung von Literaturreferenzen, die state-of-the-art Ergebnisse in der Vollständigkeit von Zitiernetzwerken erzielen. Basierend hierauf setzen wir neue Analyseformen um. (2) Wir präsentieren ein Verfahren zur Identifikation sprachübergreifender Zitate, und führen damit die bisher größte Analyse dieser Art von Zitaten durch. Durch unsere Analyse identifizieren wir Herausforderungen für die Integration nicht-Englischer Publikationen. (3) Wir entwickeln feingranulare Informationsextraktions-Methoden für

Forschungsartefakte und deren Parameter. Unsere Ansätze erzielen bessere Ergebnisse als leistungsstarke Vergleichsmethoden, und ermöglichen in ihrer Verwendung neue Formen von Analysen und Anwendungen.

Zusammengenommen adressieren unsere Beiträge zentrale Mängel existierender Methoden zur Extraktion strukturierter Daten aus wissenschaftlichen Publikationen. Durch den Einsatz unserer Methoden erzielen wir signifikante Verbesserungen im Hinblick auf Datenqualität. Für jeden unserer Ansätze demonstrieren wir dessen Umsetzbarkeit und Vorteile durch Evaluation und Anwendung auf großen Datenmengen. Die Ergebnisse unserer Arbeit haben bereits Verwendung in verschiedenen Teilen der Forschungsgemeinschaft gefunden, was deren Nutzen zusätzlich bestätigt.

Acknowledgements

I want to express my sincere gratitude to those who directly supported me in the work documented in this dissertation, as well as to everyone who is part of the wonderfully supportive environment that I enjoy.

Michael

Thank you for the support, for the collegial guidance, and for always taking the time to provide feedback.

Adam

Thank you for the encouraging words, the helpful feedback, and your always welcoming and positive attitude.

York and all my past and present colleagues at AIFB and FZI

Thank you for all the feedback, the banter, and being part of a nice and supportive work environment.

朝倉さん, 宮尾先生, Yanxia, and Prof. Kan

Thank you for welcoming me to your research group and enabling me to broaden my perspective.

Andrea, Helmut, Eric, Nora, Martha

Danke für die Unterstützung, die Zeit zusammen, und die fortwährende Gewissheit um einen Rückzugsort.

のどか

お待たせしました。応援してくれて、理解してくれて、頑張ってくれてありがとうございます。

Finally, I want to thank you, dear reader, for engaging with my work. *Thank you.*

Contents

Abstract	i
Zusammenfassung	iii
Acknowledgements	v
List of Figures	xi
List of Tables	xiii
1. Introduction	1
1.1. Scope and Motivation	1
1.2. Research Objective	2
1.3. Challenges	3
1.4. Research Gap and Tasks	5
1.5. Outline and Contributions	7
1.6. Overview of Publications	9
2. Foundations	11
2.1. Scholarly Data	11
2.1.1. Origins of Scholarly Data	12
2.2. Data Mining & Information Extraction	17
2.2.1. Data Quality	18
2.2.2. Model Evaluation	20
3. Corpus	25
3.1. Overview	25
3.2. Introduction	26
3.3. Existing Data Sets	28
3.4. Data Set Creation	30
3.4.1. Used Data Sources	30
3.4.2. Pipeline Overview	31
3.4.3. \LaTeX Parsing	32
3.4.4. Reference Resolution	33
3.4.5. Result format	35
3.5. Statistics and Key Figures	36
3.5.1. Creation Process	36
3.5.2. Resulting Data Set	37
3.6. Evaluation of Citation Data Validity and Coverage	38
3.6.1. Citation Data Validity	38
3.6.2. Citation Data Coverage	38

3.7.	Analysis of Citation Flow and Citation Contexts	41
3.7.1.	Citation Flow	41
3.7.2.	Availability of Citation Contexts	43
3.7.3.	Characteristics of Citation Contexts	43
3.8.	Conclusion	48
3.9.	Result Assessment	48
4.	Reference Coverage and Granularity	51
4.1.	Overview	51
4.2.	Reference Linking by Inter-Reference Blocking	52
4.2.1.	Introduction	52
4.2.2.	Related Work	54
4.2.3.	Approach	55
4.2.4.	Evaluation	56
4.2.5.	Discussion and Future Work	58
4.3.	Reference & Text Granularity - A Corpus Update	59
4.3.1.	Introduction	59
4.3.2.	Related Work	62
4.3.3.	Approach	62
4.3.4.	Results	64
4.3.5.	Conclusion	68
4.4.	Result Assessment	69
5.	References Across Languages	71
5.1.	Overview	71
5.2.	Introduction	72
5.3.	Related Work	74
5.3.1.	Cross-Lingual Citations in Academic Publications	75
5.3.2.	Cross-Lingual Interconnections in Other Types of Media	76
5.4.	Data Collection	76
5.4.1.	Identification of Cross-Lingual Citations	76
5.4.2.	Data Source Selection	79
5.4.3.	Data Collection	80
5.5.	Results	80
5.5.1.	Prevalence	81
5.5.2.	Usage	84
5.5.3.	Impact	92
5.6.	Discussion and Conclusion	95
5.7.	Result Assessment	96
6.	References with Usage Parameters	99
6.1.	Overview	99
6.2.	Introduction	99
6.3.	Related Work	101
6.3.1.	Fine-Tuned Models	101

6.3.2.	LLMs	102
6.4.	Hyperparameter Information Extraction	103
6.4.1.	Task Definition	103
6.4.2.	Data Set Construction	103
6.5.	Methods	105
6.5.1.	Fine-Tuned Models	106
6.5.2.	LLMs	107
6.6.	Experiments	108
6.6.1.	Fine-Tuned Models	108
6.6.2.	LLMs	111
6.7.	Discussion	112
6.8.	Conclusion	113
6.9.	Result Assessment	114
7.	Conclusion	117
7.1.	Summary	117
7.2.	Discussion	119
7.2.1.	Research Gaps	120
7.2.2.	Quality Dimensions	120
7.2.3.	Research Community	122
7.3.	Outlook	123
7.3.1.	Extensions of Our Work	123
7.3.2.	Future External Developments	124
	Bibliography of Publications	127
	Bibliography of References	129
A.	Appendix	147
A.1.	Geographic Origin of All Cited Non-English Languages	147
A.2.	Citation Intent and Sentiment Classification	147
A.3.	HyperPIE Implementation Details	148

List of Figures

1.1.	Visualisation of publications, their structured representations, and their use	1
1.2.	Schematic overview of the structural approach taken in this dissertation . . .	5
2.1.	Schematic depiction of the origins of scholarly data	13
3.1.	Schematic representation of the data set generation process	31
3.2.	Number of citing documents per cited document	37
3.3.	Number of citation contexts per reference	37
3.4.	Visualization of the citation flow in terms of documents and references from arXiv to the MAG	38
3.5.	Composition of reference section coverage for all citing documents	40
3.6.	Distribution of citing documents in terms of reference section size and their coverage in unarXive and MAG	40
3.7.	Citation flow by discipline for 15.9 million references	42
3.8.	Normalized distribution of the number of citation contexts per cited document	42
4.1.	Examples of challenging reference pairs from our evaluation that where successfully matched	53
4.2.	Schematic depiction of the use case	53
4.3.	Schematic of our data set	60
4.4.	Number of papers per year	66
4.5.	Reference density per year	66
5.1.	Schematic explanation of terminology	72
5.2.	Relative number of documents citing Russian, Chinese, Japanese, German, and French works	82
5.3.	Relative number of mathematics, physics, and computer science documents citing non-English works	83
5.4.	“Cross-linguality” of reference sections by discipline	83
5.5.	Geographic origin of cross-lingual citations to the ten most cited languages (absolute count)	85
5.6.	Geographic origin of cross-lingual citations to the ten most cited languages (relative count)	85
5.7.	Geographic origin of cross-lingual citations (local vs. English-speaking countries)	87
5.8.	Schematic explanation of an adjacent monolingual reference	88
5.9.	Comparison of citation intent distribution across arXiv categories for <i>in-text x-ling</i> and <i>in-text mono</i>	91
6.1.	Illustration of hyperparameter information in a text example alongside the extracted entities and relations	100
6.2.	Observations of initial annotation round	104
6.3.	Relation extraction with emphasis on entity candidate pair types and distance	106

6.4. Fine-tuned model evaluation (5-fold cross validation)	109
6.5. Frequency of hyperparameter mention positions in papers	110
6.6. Parsing success, format adherence, hallucinations, and scope adherence of LLM generated JSON and YAML	112
A.1. Geographic origin of cross-lingual citations (relative count)	148

List of Tables

1.1.	Overview of publications reused in this dissertation	9
1.2.	Overview of secondary publications not reused in this dissertation	10
2.1.	Comparison of Scholarly Data Sources	14
2.2.	Overview of Scholarly Data Types	16
2.3.	Overview of Scholarly Data Sets	16
2.4.	Scholarly Data Quality Criteria	18
3.1.	Overview of the proposed data set	27
3.2.	Overview of existing data sets	29
3.3.	Comparison of tools for parsing \LaTeX	32
3.4.	Confidence intervals for a sample size of 300	39
3.5.	Mismatched documents	39
3.6.	Examples of citations and their categorization into integral/non-integral as well as syntactic/non-syntactic	44
3.7.	Examples of target section specific citations	44
3.8.	Per discipline number of citations labeled integral, syntactic, imultaneously integral and syntactic, target section specific	45
3.9.	Occurrence of target section specific citations by discipline	46
4.1.	Performance of five graph weighting and graph pruning scheme combinations for meta-blocking	57
4.2.	Number of linked papers, references, and in-text citations given in the original corpus and newly created through the application of our approach	58
4.3.	Comparison of large data sets derived from paper full-texts	61
4.4.	Extension of S2ORC format	64
5.1.	Comparison of corpora	74
5.2.	References to non-Latin script languages in the automated analysis	78
5.3.	Results of manual labeling	78
5.4.	Overview of data sets	79
5.5.	Overview of data used	81
5.6.	Most prevalent languages	81
5.7.	Self-citations	84
5.8.	Class distribution and evaluation details for the model training	89
5.9.	Citation intent and sentiment classification results for cross-lingual, monolingual, and mixed in-text citations	90
5.10.	Changes in cross-lingual citations between preprints and published papers	93
5.11.	Comparison of citations received	93
6.1.	Ablation study results	109
6.2.	LLM selection	110

6.3. Prediction performance of LLM models	111
7.1. Scholarly Data Quality Criteria	117
A.1. Model configuration used for training	149

1

Introduction

Scientific publications are the *discourse medium* and *literary footprint* of academic progress. As such, they play a vital role in the everyday workings and advancement of academia. Historically, this role manifested itself physically—through ink on paper. In time, digital methods for authoring and distribution enabled more efficient ways of dissemination. Today, scientific publications are predominantly authored, distributed, consumed, and archived digitally [108]. However, the form of today’s digital publications still retains numerous and deep traces of its physical ancestry.

The historical baggage manifesting itself in the PDF files we read, share, and author, poses significant challenges to realizing the full potential of a digital record of science. While digital services in academia such as search and recommendation do exist, they are powered not by publications themselves, but rather by more structured, derivative representations of publications. The same holds for analyses of publications (see Figure 1.1). Creating such structured representations is challenging and error-prone, leading to a risk of subpar services and erroneous analyses. This dissertation represents an effort to tackle this challenge and alleviate the quality of structured representations of scientific publications.



Figure 1.1.: Visualisation of publications, their structured representations, and their use.

1.1. Scope and Motivation

The focus subject of this dissertation are *structured representations of scientific publications*, which we define as follows.

Structured Representations of Scientific Publications: Data representing the properties and contents of pieces of scientific literature in a structured way.

For the sake of brevity, we use the term “scholarly data” synonymously.

Scholarly Data: Term used synonymously to “structured representations of scientific publications” in the context of this dissertation.

The importance of scholarly data, specifically large-scale scholarly data that is of high quality, lies in its potential as a remedy to challenges brought by the digitization of academic publishing.

The digitization of academic publishing has made the transfer of new knowledge into the research community faster, thereby enabling an acceleration of scientific progress. A second driver of acceleration is the increase in research spending across the world [92, 146]. This increase in scientific progress, while first and foremost a positive development, brings with it a growing challenge for researchers to keep up with the literature. This problem is referred to as “information overload” [110]. Fortunately, the digitization of academic publishing not only lead to an increase in the rate at which research results are being published. It also marks the inception of scholarly data, and thereby enabled search technology and analytics to operate on large collections of digitally archived publications. In this way, the existence of digital representations of publications also provides, to some degree, a remedy for information overload. Efficient search and recommendation services, for example, can aid researchers in navigating the deluge of publications they are faced with. Similarly, decision processes in academia, such as the evaluation of institutions or researchers, is enabled by scholarly data through performance indicators. In this way, it provides a means for hiring and funding decisions. In other words, scholarly data is a vital resource for decision-making in academia on the individual as well as organizational level.

The quality of decisions made based on scholarly data, naturally, hinges on the quality of the scholarly data itself. If, for example, citations are missing in the data of a search engine, this can cause researchers to overlook relevant related work, or a funding body to under-evaluate an institution. Recalling that the creation of scholarly data is a challenging and error-prone process, it stands to reason that efforts to improve scholarly data quality are a worthwhile endeavor. Based on these considerations, the overarching objective pursued in this dissertation is the development of methods for creating high-quality scholarly data.

1.2. Research Objective

To define our research objective, we need to take a closer look at what “creating high-quality scholarly data” entails. As briefly laid out above, scholarly data is a derivative product of publications. This is due to the fact that, for the time being, scientific publications are written by humans for humans. Because the product of this process—i.e., scientific literature in the form humans consume it in—is not machine-readable *as is*, scholarly data is created as a derivate based on publications. To make this creation process feasible on a large scale, it has to be automated.

Accordingly, we can define our research objective as follows.



Research Objective

Develop an automated process that takes as input scientific publications, and produces as output a high-quality, machine-readable derivative representation of the publications.

This leaves the question of what “high-quality” entails. We discuss this aspect in detail in Chapter 2. In short, data quality depends on the intended use and therefore, generally speaking, means *fitness for use*. Quantitatively, data quality is most commonly assessed along the following five dimensions [81].



Data Quality Dimensions

(1) relevance, (2) accuracy, (3) timeliness, (4) comparability, (5) completeness

The five dimensions can briefly be described as follows.

1. **Relevance** measures the extent to which the data is connected to its intended use.
2. **Accuracy** is concerned with how error-free the data is.
3. **Timeliness** expresses to which degree the data is current enough.
4. **Comparability** quantifies how well the data is comparable to other data, e.g. due to the existence of common identifiers.
5. **Completeness** is a measure for the rate of missing records, and the rate of missing data elements within records.

We elaborate on this in Section 2.2.1, where we derive specific quality criteria for scholarly data across these dimensions, grounded in considerations of how scholarly data is used.

1.3. Challenges

Developing an automated process to generate high-quality scholarly data as laid out above is challenging for several reasons. In the following, we describe the challenges connected to each of the five data quality dimensions.

Relevant scholarly data is challenging to attain especially for two reasons. First, common use cases such as bibliographic analyses require a large **volume of data** (often millions of documents) in order to yield meaningful results. This leads to challenges in terms of processing efficiency, robustness, and fault tolerance. Second, relevant content of scientific literature is of **multiple modalities**, such as natural language text, mathematical notation, figures, etc. This necessitates the use of either a combination of multiple, specialized processing approaches, or a highly generalizable processing approach.

Accurate scholarly data proves difficult to obtain, because, as described earlier, scientific publications are created for human consumption, and therefore presuppose visual

parsing by a system with relevant background knowledge, which leads to a challenge of **information sparsity**. In other words, it is difficult to bridge the gap between a human-oriented medium and machine-readable data. Furthermore, the contents of scientific publications pose a challenge for accuracy due to their **specialized content**. Because scientific publications address highly specialized topics, they are likely to contain both specialized terminology, as well as specialized notation. Both of these aspects result in further challenges in processing their natural language contents.

Timely scholarly data can be difficult to realize especially in areas of research that have a high **publishing rate**. This is because integrating the data derived from newly published work into an existing corpus can necessitate (re-)processing steps across of the entire corpus. This in turn leads to challenges in processing efficiency.

Comparable scholarly data poses a challenge, because of an **ambiguity of labels**, which is amplified by the **specialized content** of scientific publications. For example, approaches and resources in computer science are often referred to by ambiguous names or acronyms, such as the language model “BERT” [50] or the “iris” data set [62], which are identical to a common name and noun respectively. Similarly, literature references can be ambiguous when no unique identifier such as a digital object identifier (DOI) is given. In each of these cases, it is challenging to accurately identify what is being referred to, which is important for the creation and use of structured data representations.

Complete scholarly data is challenging to obtain for multiple reasons. First, similar to the dimension of relevance, the **multiple modalities** contained in scientific literature render the task of completely representing publications’ content in data difficult. Second, literature references turn the target of data processing into an **unbounded domain**. This is because, even for a constrained set of publications, it is practically infeasible for a process generating scholarly data to guarantee access to everything possibly referenced by these publications. Lastly, there is a challenge of **multi-linguality**. While English currently is the de facto academic lingua franca [140], science is a global endeavor, meaning that scientific publications are written in various languages. Accordingly, the creation of scholarly data can entail the challenges that come with processing multi-lingual text.

To various degrees, some of the aspects mentioned above apply to several other tasks in natural language processing (NLP), such as an ambiguity of labels. However, in their entirety, these challenges set the creation of scholarly data apart from other areas. For example, approaches concerned with news articles contain less specialized content, and in the case of websites the source material is more structured in nature. Rather close to scientific publications in terms of challenges is the processing of patents. However, patents also contain legal jargon, which is generally not the case for scientific publications.¹

¹ The similarity between scientific publications and patents regarding the nature of their content, purpose, and requirements for automated processing, are likely the reason that some platforms, such as Google Scholar, handle both within a single system. The search interface at <https://scholar.google.com/> [last accessed: 2023-11-22], for example, includes a filter option “include patents”.

		Research Objective			
		Base	Research Gap		
		Research Task 1	Research Task 2	Research Task 3	Research Task 4
Data Quality	D ₅	✓ X	✓ X	✓ X	✓ X
	D ₄	✓ X	✓ X	✓ X	✓ X
	D ₃	✓ X	✓ X	✓ X	✓ X
	D ₂	✓ X	✓ X	✓ X	✓ X
	D ₁	✓ X	✓ X	✓ X	✓ X

Figure 1.2.: Schematic overview of the structural approach taken in this dissertation. The overall objective of improving scholarly data quality is quantified across data quality dimensions (rows). The approach taken towards this goal is focussed by identifying the research gap and deriving research tasks. Including an initial step of establishing a base upon which the subsequent research is built, four research tasks are identified (columns). The cell contents “✓ | X” indicate an evaluation of the research task results for the respective data quality dimension.

1.4. Research Gap and Tasks

This dissertation is not the first research endeavor with the objective to develop methods for the creation of high-quality scholarly data. Prior approaches exist and, accordingly, have grappled with the challenges laid out in the previous section. However, there are three key areas of particular importance in which we see shortcomings in existing work. These allow us to focus our efforts of improving scholarly data quality and, as a whole, make up the research gap addressed by this dissertation. We describe the three areas in the following and, based on that, formulate our research tasks. Figure 1.2 provides a structural overview of how the research gap and research tasks relate to the previously defined data quality dimensions.

1. **Citation Network** Several of the most prevalent use cases for scholarly data hinge on the interlinking of publications through citations. This includes use cases such as bibliometric analyses, scientometrics, as well as the study of citation networks as a type of graph (e.g. in graph neural network evaluations). Despite this importance, not much focus seems to be put on the quality of citation networks in scholarly data. The much used data set CiteSeerX [199, 198, 152], for example, takes the approach of clustering references and publications, but there is no assessment of the citation network’s completeness—i.e., what proportion of references of a paper can successfully be linked to the cited document. The more recently introduced data set S2ORC [122]—more specifically, its \LaTeX subset—provides an investigation into the proportion of successfully matched references, but only achieves to successfully match 31.1% of references. It stands to reason that data, in which over two thirds of the records are missing a key element, or where the proportion of records missing a key element is unknown, is of insufficient quality. Improvements regarding citation

networks in scholarly data are presented in Chapters 3 and 4. The results presented there entail improvements across all five data quality dimensions.

2. **Anglocentrism** Science is a global and therefore inherently multi-lingual endeavor. Accordingly, capturing the state and progress of science in data, necessitates including publications written in various languages. In several areas of NLP, there has been a growing awareness and consideration of non-English language content in the last years [49, 206, 162]. However, in research concerned with scientific publications, as well as in available data sets covering scientific publications, there still is a major lack of coverage of non-English documents [189, 121, 137, 142, 127]. Accordingly, research results are of limited validity due to an insufficiency of the underlying data. Improvements regarding the inclusion and analysis of non-English publications are presented in Chapter 5. The presented work entails improvements in the data quality dimensions relevance and comparability.
3. **Research Artifacts** To an increasing degree, research is driven by curated data sets and algorithmic processing techniques, such as machine learning methods and models (“research artifacts”). This development can, for example, be observed in the field of NLP, which has undergone a shift towards “rapid discovery science”, characterized by a high consensus on research topics, methods, and technologies [95]. Further signs of the growing importance of research artifacts are, for example, the launch of both Google Dataset Search² and Papers With Code³ in 2018, as well as the gradual adoption of data citations [103]. Despite these developments, efforts to include structured representations of research artifacts in scholarly data are limited. Some work in this direction exists, most notably SciERC [124] and SciREX [87], but it only covers shallow representations of research artifacts. We argue that shallow representations are insufficient, as they do not provide the necessary granularity for several beneficial applications, such as faceted search and automated reproducibility. Improvements regarding a more fine-granular coverage of research artifacts are presented in Chapter 6. Similar to anglocentrism, the work presented entails improvements in the data quality dimensions relevance and comparability.

In addressing above limitations, we alleviate the state of scholarly data. This improvement is also reflected in the data quality dimensions introduced in Section 1.2, as mentioned for each of the three points above. Details regarding the improvements are tracked along the dissertation, and are summarized at the end of each chapter. The steps to address the identified research gap are structured into four research tasks, described in the following.

Research Tasks Based on the research gap identified, we formulate the following four research tasks as sub-points to our overarching research objective set in Section 1.2.

² See <https://datasetsearch.research.google.com/> [last accessed: 2023-11-24].

³ See <https://paperswithcode.com/> [last accessed: 2023-11-24].

- ❖ **RT1:** *Base Methodology* - establish a base methodology for generating a large-scale, high-quality scholarly data set, that is on par with or improving upon existing data sets.
- ❖ **RT2:** *Citation Network Completeness* - develop a method to link literature references, that is able to link more references than are linked in existing data sets, while not compromising on link correctness or processing efficiency.
- ❖ **RT3:** *Inclusion of Non-English Publications* - find and implement an approach to include non-English publications into a large-scale, high-quality scholarly data set.
- ❖ **RT4:** *Fine-grained Research Artifact Representations* - develop a method to extract fine-grained information on research artifacts from text in scientific publications.

1.5. Outline and Contributions

The contributions made across the works tackling the challenges and research gaps outlined above, represent the main part of this dissertation. Together with the following and the final chapter, they make up the remainder of the document, which is structured as follows.

- Chapter 2 - **Foundations**
- Chapter 3 - **Corpus**
- Chapter 4 - **Reference Coverage and Granularity**
- Chapter 5 - **References Across Languages**
- Chapter 6 - **References with Usage Parameters**
- Chapter 7 - **Conclusion**

In Chapter 2, Foundations, we introduce overarching and foundational concepts related to scholarly data as well as data mining and information extraction. The provided information is conducive to understanding (1) decisions made in the system design and method development of the approaches discussed later on, and (2) the quantification of the research goals and achieved results.


Chapters 3 to 6 make up the main contributions of the work presented in this dissertation. Specifically, these are:

Chapter 3 - **Corpus**

- Contribution: *Linked Document Scholarly Data Corpus Creation from L^AT_EX*
- Addresses: ❖ **RT1**, ❖ **RT2**
- Improves: relevance, accuracy, timeliness, comparability, and completeness


With *unarXive* we present in Chapter 3 a methodology for creating a large-scale corpus of linked, full-text documents from \LaTeX source files, which we apply to all of arXiv.org. At the time of publication, it was the first corpus of linked publications with extensive coverage in physics, mathematics, and computer science. By creating the corpus from \LaTeX source files, it is less noisy than related work created from PDFs, such as CiteSeerX [199]. The creation method furthermore includes a highly accurate reference matching procedure achieving a state-of-the-art matching success rate. This contribution lays the foundation for the subsequent research conducted for the dissertation, and primarily addresses the research gap *citation network*. The work presented achieves improvements across all five data quality dimensions.

Chapter 4 - Reference Coverage and Granularity

- Contribution: *Inter-Reference Blocking and Fine-Granular Text Representation*
- Addresses:  **RT2**
- Improves: relevance, timeliness, comparability, and completeness


Building upon *unarXive*, we present in Chapter 4 advancements in two areas. First, regarding the citation network, we develop an inter-reference blocking and matching method that significantly increases matched references as well as bibliographic couplings, and achieve a new state-of-the-art matching success rate. Second, we present an improved conversion method for \LaTeX source files leading to fine-granularly structured document representations. This contribution addresses the research gap *citation network*, and furthermore lays the foundation for the research presented in Chapter 6 (see below). In total, the work presented achieves data quality improvements in terms of relevance, timeliness, comparability, and completeness.

Chapter 5 - References Across Languages

- Contribution: *Detection and Large-Scale Analysis of Cross-Lingual Citations*
- Addresses:  **RT3**
- Improves: comparability and completeness

Using the *unarXive* corpus, we present in Chapter 5 a method to reliably identify cross-lingual citations in English publications. Based on this, we conduct the so far largest analysis of this type of citation. Where previous studies of comparable setting only looked at hundreds of documents, our study includes over one million publications. Analyzing cross-lingual citations' prevalence, usage, and impact, we identify trends over time as well as challenges. This contribution addresses the research gap *anglocentrism*. The work results in data quality improvements in terms of relevance and comparability.

Chapter 6 - References with Usage Parameters

- Contribution: *Information Extraction for Hyperparameter Information*
- Addresses:  **RT4**
- Improves: comparability and completeness

To further improve the granularity of the full-text representations in *unarXive*, we develop in Chapter 6 information extraction methods for research artifacts and their usage parameters. In doing so, we enable the study of parameter use and reporting patterns across time and scientific disciplines. The extracted information furthermore bears potential for use in automated reproduction. The developed methods achieve an improvement over strong baselines. This contribution addresses the research gap *research artifacts*. Regarding data quality, the work represents improvements in the data quality dimensions relevance and comparability.

The dissertation concludes in Chapter 7, with an overarching discussion of the results attained, impact of the work so far, and an outlook.

1.6. Overview of Publications

The contributions in this dissertation have been published in peer-reviewed international conferences and journals. Table 1.1 gives an overview of the publications and the chapters they make up. Venue ranks are taken from Core⁴ in the case of conferences and from SJR⁵ in the case of journals.⁶ For all publications in Table 1.1, the author of this dissertation is the first and corresponding author. Detailed author contributions according to the Contributor Roles Taxonomy⁷ are listed at the end of the respective chapter.

Table 1.1.: Overview of publications reused in this dissertation.

Chapter	Venue	Rank ^a	Type	Year	Length	Ref.
3	Scientometrics	SJR Q1	Journal	2020	Full	[1]
4	ULITE@JCDL	Core A*	Workshop	2022	Full	[2]
	JCDL	Core A*	Conference	2023	Short	[3]
5	ICADL	Core A	Conference	2020	Full	[4]
	IJDL	SJR Q2	Journal	2022	Full	[5]
6	ECIR	Core A	Conference	2024	Full	[6]

^a Venue rank in publication year (or closest prior). For workshops, the rank of the hosting conference is shown.

Additional publications (co-)authored leading up to and during the research period, which are not a direct part of this dissertation but nevertheless informed the overall research trajectory, are listed in Table 1.2.

⁴ See <http://portal.core.edu.au/conf-ranks/> [last accessed: 2023-10-12].

⁵ See <https://www.scimagojr.com/> [last accessed: 2023-10-12].

⁶ The ranks shown are the rating for the respective publication year, or, if not available, the most up-to-date prior ranking. For workshops, the rank of the conference at which the workshop is hosted is shown.

⁷ See <https://credit.niso.org/> [last accessed: 2023-10-12].

Table 1.2.: Overview of secondary publications not reused in this dissertation.

Venue	Rank ^a	Type	Year	Length	Pos. ^b	Ref.
BIR@ECIR	Core A	Workshop	2019	Full	1 of 2	[7]
ECIR	Core A	Conference	2020	Full	1 of 3	[8]
SDP@NAACL	Core A	Workshop	2021	Short	3 of 4	[9]
SDU@AAAI	Core A*	Workshop	2022	Full	2 of 3	[10]
BIR@ECIR	Core A	Workshop	2022	Full	4 of 5	[11]
JCDL	Core A*	Conference	2022	Full	3 of 3	[12]
JCDL	Core A*	Conference	2023	Short	1 of 3	[13]

^a Venue rank in publication year (or closest prior). For workshops, the rank of the hosting conference is shown.

^b Author position.

Especially [7] and [8], which constitute the results of the master’s thesis preceding the doctoral research period, paved the way for this dissertation.

2

Foundations

This chapter provides information on overarching as well as foundational concepts relevant to the work presented in this dissertation. Specifically, we cover two areas.

1. Scholarly Data

First, we give an overview of the academic publication ecosystem and its relation to the landscape of scholarly data. Understanding the parts involved and relations between them is helpful for understanding decisions made in the system design and method development of the approaches presented later on. Based on the overview, we additionally highlight the current state of the art.

2. Data Mining & Information Extraction

Second, we present essential concepts from the areas of data mining and information extraction. These are needed for the quantification of the research goals as well as the results that are presented later on.

Explanations of concepts that are specific to the work presented in individual chapters, as well as a focussed view on state-of-the-art approaches in the respective areas, are provided jointly with the approaches in Chapters 3 – 6.

2.1. Scholarly Data

The term “scholarly data” is used in this dissertation to refer to data that represents academic publications. It can coarsely be divided into data directly reflecting the *content* of publications, and metadata, which gives information *about* publications. As such, scholarly data is the basis for essentially everything that relies on automated processing of publications. The following are three key examples.

1. **Digital services** in academia, such as search (e.g. Google Scholar,¹ Semantic Scholar²), recommendation (e.g. Academia.edu,³ CORE Recommender⁴), and aggregation platforms (e.g. Papers With Code,⁵ Scopus⁶).
2. **Analyses**, such as bibliometric analyses across time, geographic regions, or institutions, as well as trend analyses and investigations into specific phenomena like citation inequity.
3. **Model development**, such as the training and evaluation of transformer based large language models (LLMs), as well as task specific models (e.g. for recommender systems, impact prediction, or information extraction).

Because scholarly data is only a secondary product to the actual publications themselves, it is necessary to consider how the data comes into being.

2.1.1. Origins of Scholarly Data

Figure 2.1 schematically shows the path of a publication from authorship to distribution, together with different stages from which scholarly data can emerge. It is essential to note, that academic publications are, historically and at the time of writing still, primarily written by humans with human readership in mind. As such, publications are primarily a visual medium, optimized for parsing by human vision and intelligence. Scholarly data, however, is intended for automated processing and therefore benefits, for example, from strict syntactic rules and no reliance on assumed background knowledge. As a consequence, the creation of scholarly data requires bridging the gap between the existing visual presentation and desired structural derivate.

Specifically, this means that information, which is not made explicit in publications, needs to be retroactively added. For example, whether a piece of text “[1,3]” in a publication is expressing an interval of real numbers, or a citation for references 1 and 3. Because it would be impractical to require researchers to produce a detailed set of annotations in addition to all their publications,^{7,8} the retroactive adding of information needs to be done automatically, i.e., by means of information extraction. As shown in Figure 2.1, the information extraction can happen at any given stage of publication. At each stage, the nature of the available data and information is different. Accordingly, there are different benefits and challenges in each case. In the following, we will discuss these different types

¹ See <https://scholar.google.com/> [last accessed: 2023-11-06].

² See <https://www.semanticscholar.org/> [last accessed: 2023-11-06].

³ See <https://www.academia.edu/> [last accessed: 2023-11-06].

⁴ See <https://core.ac.uk/services/recommender> [last accessed: 2023-11-06].

⁵ See <https://paperswithcode.com/> [last accessed: 2023-11-06].

⁶ See <https://www.scopus.com/> [last accessed: 2023-11-06].

⁷ This can be seen as a case of the “authoring problem” challenging the semantic web community [100].

⁸ An exception to this is basic metadata such as title, authors, and abstract, which are commonly requested to be filled into a form in plain text during the submission process of a manuscript (see Figure 2.1 top left).

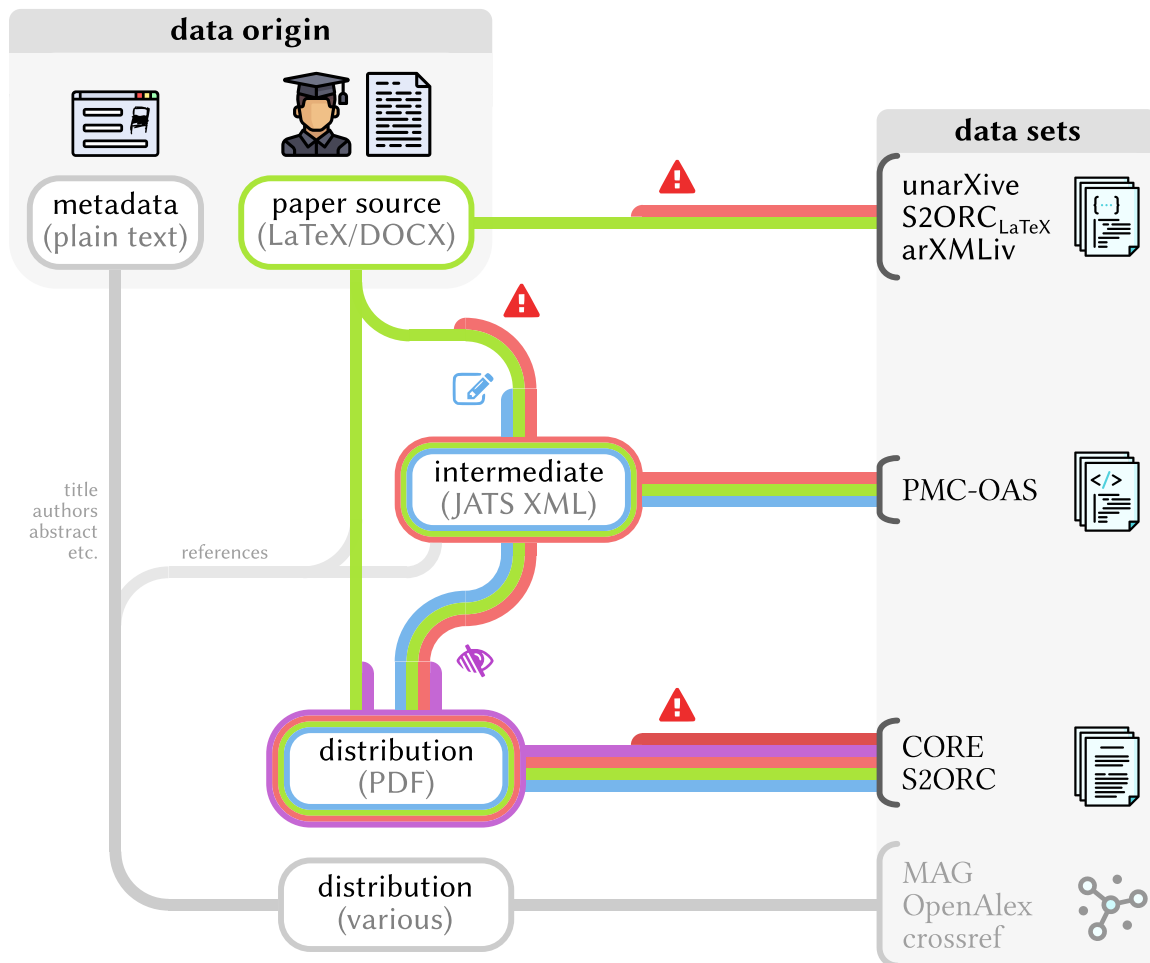


Figure 2.1.: Schematic depiction of the origins of scholarly data. Starting from the data origins in the top left and ending in scholarly data sets on the right. In between, each processing step is indicated by a symbol and an additional colored line. Symbols indicate (1) error-prone automated processing to extract structural and semantic information, (2) a manual check, and (3) loss of structural information due to transformation into a visual format. Physical paper sources, which require optical character recognition (OCR) of scans for the creation of scholarly data, are not depicted.

of data from an information extraction perspective, beginning with \LaTeX as it is the most relevant to the presented work.

2.1.1.1. Types of Data Sources

Viable input formats for the creation of scholarly data differ in terms of the contained information, challenges for information extraction, availability, and use. These are discussed in the following, with Table 2.1 providing an overview.

\LaTeX is described as “a system for typesetting documents” [109], or a “widely used language for describing the logical structure of [...] documents” [134]. At its core, \LaTeX provides

Table 2.1.: Comparison of Scholarly Data Sources.

Source	Doc Structure	Semantic	Open Availability	Disciplines
\LaTeX	✓	partial ^a	✓ (arXiv)	physics, math, CS
Word	✓	partial ^b	× (publisher internal)	all ^c
JATS	✓	✓	✓ (PMC OAS)	biomedical
PDF	×	×	✓ (abundant)	all

^a Semantic to some degree by use of named macros (e.g. `\title{}` or `\section{}`), but usage not enforced by format.

^b Semantic to the extent that the DOCX (Office Open XML) schema describes the document structure, but dependent on authors' usage of respective Word features.

^c Primarily humanities; STEM fields only to a lesser degree.

functionalities to control the visual presentation of a document. These functionalities are combined by macros, which offer authors the means to structurally and semantically describe a document (e.g. `\title{}` or `\section{}`). This structural and semantic information—which gets lost when the \LaTeX source is compiled to PDF—is immensely beneficial for information extraction. However, \LaTeX documents are usually a mixture structural and presentation description, which introduces challenges [174]. While there have been efforts to establish \LaTeX extensions for more rigorous semantic annotation in the document source [104, 73, 30], these have not been widely adapted so far. There are, however, efforts by The \LaTeX Project⁹ itself to support semantic annotation natively in the future [135, 134].

Regarding at the availability of \LaTeX sources, arXiv.org¹⁰ provides open access to over 2 million papers uploaded by their authors. With its origin in physics in the 1990s [61, 69], gradual adoption in mathematics in the early 2000s, and rapid growth in computer science since the 2010s [3], it now covers significant portions of the scientific literature in aforementioned three disciplines. Given the benefits for information extraction, \LaTeX sources from arXiv have been used for generating scholarly data on a small scale since at least 1998 [143]. Large-scale efforts started with a focus on mathematics in 2008 [174]. Complete conversions of the papers on arXiv.org into a scholarly data corpus including a citation network are comparatively new development with the unarXive corpus [1] presented in this dissertation, as well as S2ORC [122]. Besides arXiv.org, we are not aware of other significant sources of scientific literature in \LaTeX format. Related platforms such as bioRxiv¹¹ and HAL¹² do only offer documents in PDF format and not as \LaTeX sources.

⁹ See <https://www.latex-project.org/> [last accessed: 2023-11-08].

¹⁰ See <https://arxiv.org/> [last accessed: 2023-11-08].

¹¹ See <https://www.biorxiv.org/> [last accessed: 2024-02-03].

¹² See <https://hal.science/> [last accessed: 2024-02-03].

Word documents (DOCX files¹³) are the second major data format commonly accepted by publishers for submitting manuscripts [93]. Contrary to the \LaTeX approach of compilation from source files, documents are edited in an interactive “What you see is what you get” (WYSIWYG) editor.¹⁴ While DOCX files are essentially ZIP compressed XML files, and contain more explicit information than PDF derivatives, there appear to be no open repositories similar to arXiv.org that provide large quantities of papers’ Word source files. These are only available to the publishers receiving manuscript submissions in DOCX format.

Journal Article Tag Suite (JATS) is a standardized markup format for scientific publications based on XML [84]. As depicted in Figure 2.1, JATS files are not directly created by researchers, but are rather an intermediate format used by publishers, from which they derive different presentation formats of publications, such as PDF and HTML. While JATS files provide semantically richer information than \LaTeX or Word source files, their generation can only partially be automated and requires human oversight. Regarding availability, the PubMed Central Open Access Subset¹⁵ (PMC OAS) provides over 3 million publications from the biomedical and life sciences domain in JATS XML format. While the JATS files of the PMC OAS have been used to generate a corpus of linked publications in the past [70], more up-to-date and widely used corpora have only used the PDF versions of the contained documents [122].

PDF is the most common distribution format for academic publications [93] and accordingly, the largest open document collections are PDF files, such as CORE [154] with over 100 million documents. The PDF format does provide optional functionalities to describe the logical structure of documents in addition to the visual presentation [86]. However, such annotation is not an established practice in academic publishing and, accordingly, information extraction methods have to resort to heuristic approaches based on the visually presented information only [123, 144, 58]. This makes PDF a more error-prone data source than aforementioned source formats [24].

2.1.1.2. Types of Scholarly Data

Conceptually, scholarly data can be divided into two overarching categories: *metadata* and *document collections* [144]. As a third category, we consider *linked document collections*, which combine features of aforementioned two. We briefly introduce each of the three types in the following. An overview by type can be found in Table 2.2, while a summary of representative data sets for each type is shown in Table 2.3.

¹³ See <https://www.loc.gov/preservation/digital/formats/fdd/fdd000397.shtml> [last accessed: 2023-11-08].

¹⁴ See <https://www.microsoft.com/microsoft-365/word> [last accessed: 2023-11-08].

¹⁵ See <https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/> [last accessed: 2023-11-06].

Table 2.2.: Overview of Scholarly Data Types.

Type	Contents	Size ^a	Examples
Metadata	Title, author, citations, etc.	10 ⁸	OpenAlex, ORKG
Document Collections	Full-text	10 ⁸	CORE, arXMLiv
Linked Doc. Collections	Full-text + citation network	10 ⁷	unarXive, S2ORC

^a Order of publications covered by largest representatives of each type

Table 2.3.: Overview of Scholarly Data Sets.

Type	Data Set	# Docs	Main Sources
Metadata	OpenAlex	> 200,000,000	PubMed, arXiv, publishers
	ORKG	> 28,000	Manual/semi-automated input
Document Collection	CORE	> 140,000,000	Open access repositories
	arXMLiv	> 1,600,000	arXiv
Linked Doc. Collection	S2ORC (full)	> 12,000,000	PubMed, arXiv, publishers
	CiteSeerX	> 10,000,000	PubMed, publishers
	unarXive	> 1,900,000	arXiv
	S2ORC (L ^A T _E X)	> 1,500,000	arXiv
	SciXGen	> 200,000	arXiv

Metadata provides information *about* publications, rather than reflecting their full-text content. The data partially originates in already structured form provided by authors, as it is queried by publishers during manuscript submission. This includes but is not limited to the title, authors, and the abstract. Not queried in dedicated input forms during manuscript submissions are bibliographic references. They are, however, often included in metadata sets. To achieve, this it is necessary to extract the reference information from the document submitted by an author (e.g. L^AT_EX or Word sources). Regarding accessibility, title and author information is generally shared freely. Abstracts and references may also be openly acceptable, but this is not always the case, as evidenced by the existence of the Initiative for Open Abstracts¹⁶ and the Initiative for Open Citations.¹⁷

Two examples of scholarly metadata sets are the following. (1) OpenAlex [156] contains data on over 200 M publications, their authors, affiliations, citation links, etc. Its data sources include PubMed, arXiv, academic publishers, and various institutional repositories. (2) The Open Research Knowledge Graph (ORKG) [175, 21] provides information on over 28 k publications,¹⁸ their contributions, research problems addressed, etc. For its data, the ORKG is largely reliant on manual or semi-automated data entry.

¹⁶ See <https://i4oa.org/> [last accessed: 2023-11-09].

¹⁷ See <https://i4oc.org/> [last accessed: 2023-11-09].

¹⁸ 28.020 entities of type <http://orkg.org/orkg/class/Paper> as of 2023-11-09. Determined via the ORKG SPARQL endpoint at <https://www.orkg.org/sparql> [last accessed: 2023-11-09].

Document Collections provide access to publications' full-text content. Because of this, they are reliant on open access publications as a data source. Furthermore, the documents either need to be licensed in a way that allows (re-)distribution, or access to the collection itself needs to be restricted accordingly. Because there are document sources in different formats (see Figure 2.1), document collections derived from them take different forms. While some are primarily aggregates of openly accessible PDFs (and therefore can also be seen as a data source), such as CORE [154], others are the result of an involved generation process. An example for the latter is arXMLiv [68], which is an XML conversion of the papers on arXiv, comprising 1.6 M documents in its most recent version.

In order to use a document collection for applications involving citation information, it has to be jointly used with a suitable set of citation metadata. This requires (1) the citation metadata to cover the documents within the collection, and (2) either the availability of identifiers (such as DOIs) or some matching procedure.

Linked Document Collections are document collections that, in addition to the document contents, include a citation network. This means, that in addition to the requirements for creating a document collection, their generation involves the additional step of linking the references in the documents' reference sections. Commonly, this furthermore involves linking in-text citations to their respective references, to provide for more fine-grained citation information.

Representatives of linked document collections are CiteSeerX [199, 198, 152], S2ORC [122], unarXive [1], and SciXGen [38]. unarXive and SciXGen are generated from \LaTeX sources on arXiv, and comprise about 2 M and 200 k documents respectively. unarXive, utilizing all of arXiv, covers physics, mathematics, and computer science publications. SciXGen, on the other hand, only includes computer science. CiteSeerX and S2ORC are both generated from PDF files of various sources, such as PubMed and publishers' websites, and contain over 10 M and 12 M documents respectively. The publications covered are from various fields, including medicine and biology, physics, mathematics, and computer science.

2.2. Data Mining & Information Extraction

The work presented in this dissertation aims to improve the quality of scholarly data, and involves both data mining, the process of "discovering useful patterns and trends in large data sets" [111], and information extraction, the process of "converting unstructured text into a structured representation" [16]. In the following, we therefore provide background information regarding (1) data quality, and (2) the evaluation of information extraction models. Based on the considerations and concepts introduced, the goals and results of the work presented in the following chapters can be quantitatively assessed.

Table 2.4.: Scholarly Data Quality Criteria.

Dimension	Focus ^a	Specific Criterion and Short Description	
Relevance	CN	\mathbf{Rel}_{CN}	Representative coverage of publications in area of study
	SDR	$\mathbf{Rel}_{\text{SDR}}$	Inclusion of relevant content types (text, math, etc.)
Accuracy	CN	\mathbf{Acc}_{CN}	Correctly linked references
	SDR	$\mathbf{Acc}_{\text{SDR}}$	Noise-free full-text content
Timeliness	both ^b	$\mathbf{Tim}_{\text{C/S}}$	Coverage of recent publications
Comparability	CN	\mathbf{Coy}_{CN}	Use of established doc. identifiers (DOI, PMID, etc.)
	SDR	$\mathbf{Coy}_{\text{SDR}}$	Fine-granular, specifically typed content representation
Completeness	CN	\mathbf{Cos}_{CN}	All references in publications successfully linked
	SDR	$\mathbf{Cos}_{\text{SDR}}$	No sections or content missing (appendices, math, etc.)

^a Focus area of use (CN = Citation Network, SDR = Structured Document Representation)

^b Bundled together because the timeliness of a publication's content (SDR focus), is bound to the timeliness of the of publication itself (CN focus). This is because neither part of the document changes independently of the other.

2.2.1. Data Quality

Data quality is commonly understood as being context specific. In a general way, data quality is accordingly defined as “fitness for use” [176, 94]. Meaningful metrics for data quality in the context of scholarly data, can therefore be derived by considering scholarly data use. Because existing literature on scholarly data quality only considers narrow use cases and offers no systematic consideration of the topic [177, 112], we present a structured, more encompassing perspective in the following.

2.2.1.1. Data Quality in the Context of Scholarly Data

Historically, the dominant use of scholarly data is found in bibliometric and scientometric analyses [66, 133], such as impact or performance analysis and trend detection. The focus of use in that case lies on the citation network. More recently, with the continuing rise of open access publishing and advances in NLP, additional areas of significant use have been established. These are training ML models [74], and analyses of publications' full-text content [95, 106]. While the citation network remains a key focus area for scholarly data use [196], these recent usage patterns additionally put importance on structured representations of the publication content.

Quantitatively, data quality is most commonly assessed along the five dimensions (1) relevance, (2) accuracy, (3) timeliness, (4) comparability, and (5) completeness [81]. Based on these dimensions and aforementioned focus areas of scholarly data use, quality criteria for scholarly data can be derived, as shown in Table 2.4. In the following, we discuss the five dimensions with regard to scholarly data, as well as the quality criteria derived from the focus areas of use.

1. **Relevance** is a quality dimension heavily dependent on the data’s intended use. As scholarly data is used to gain insight on particular aspects of academia (e.g. a certain field, practice, institution or individual), the data needs to cover that aspect to a sufficient degree in order to be relevant. For citation based analyses, this means a representative portion of publications from the studied disciplines, time span, languages, etc. has to be covered (\mathbf{Rel}_{CN}). On the level of document representations, it is important that relevant content types are included. For example, for a comparison of mathematical formula usage in publications, the data has to contain publications’ full-text *including* pieces and sections of mathematical notation (\mathbf{Rel}_{SDR}).
2. **Accuracy** regarding the citation network means, that references should be correctly linked to the publications they are referencing (\mathbf{Acc}_{CN}). For document representations, accuracy entails that text content should be free of noise (\mathbf{Acc}_{SDR}). As described in the previous sections, both of these are non-trivial due to the sources scholarly data relies on. Regarding the question of how accurate scholarly data has to be for meaningful usage, previous research proposes that for “local analyses” (e.g. node degrees in a citation network) 80% correct data can already allow for reasonable insight, but for “global analyses” (e.g. rankings) at least 90% correctness should be aimed for [177]. Others have argued that quality requirements in scientometrics are high enough to warrant the cost of requiring a “human in the loop” approach to data generation [112].
3. **Timeliness** with regard to scholarly data is crucial if the underlying reality of the research object changes quickly. In cases like training a model for parsing bibliographic references, recent data is likely not needed, as citation styles do not change rapidly. However, for tasks like paper recommendation, everything published after the recommender model’s training data simply cannot be recommended. Accordingly, the coverage of recent publications is desirable. For timeliness, there is no distinction between the citation network and document representations, because neither part of the document changes independently of the other ($\mathbf{Tim}_{S/R}$).
4. **Comparability** as a quality criterion is of importance to enable the combined use of data sets. In other words, items in the data set should be clearly identified. In the case of scholarly data, this calls for determining the persistent identifiers (e.g. DOI, ORCID, and ROR [128]) of items of interest. For the citation network specifically, this means providing document identifiers (\mathbf{Coy}_{CN}). Regarding a structured document representation, comparability of contents can be facilitated by providing a fine-grained, typed content representation—for example, by representing a publications’ text structured into sections, subsections, etc. (\mathbf{Coy}_{SDR}).
5. **Completeness** of data means that no records are missing, *and* that records are not missing any attributes. Regarding scholarly data, there are practical boundaries to simply aiming for including everything ever published (e.g. closed access publications and distribution rights). It is reasonable, however, to strive for a complete citation network—that is, that all entries in the reference section of a given publication are successfully linked (\mathbf{Cos}_{CN}). For a structured document representation, completeness

means that no parts of the content or types of content are missing, like appendices, mathematical notation, etc. (Cos_{SDR}).

Based on these considerations, and the resulting quality criteria presented in Table 2.4, a quantitative view on the goal of the presented work (enabling higher quality scholarly data) and achieved results is possible. At the end of each chapter, we present a summary box as the one shown below, accompanied by a brief discussion of each point.

Scholarly Data Quality Contributions - [Example]	
Crit.	Contribution
Rel_{CN}	Improvement of relevance, by adding ... to the citation network.
Rel_{SDR}	Improvement of relevance, by adding ... to the structured document representation.
...	...
Cos_{SDR}	Improvement of completeness of document representation, by

In the following section, we present the tools to quantitatively assess the information extraction models that enable the improved scholarly data.

2.2.2. Model Evaluation

Model evaluation guides the development and assessment of approaches to computational problems. For example, when developing a machine learning approach to a classification problem, it is of interest to compare the approach to existing methods. In essence, a model evaluation results in a performance estimate. This is because the model performance can not be known in advance for every possible input, so only an estimate based on a set of realistic inputs can be attained. To get a performance estimate, the model is applied on inputs, for which the desired output is already known—a *ground truth*. The model outputs are then compared with the desired outputs by means of *evaluation metrics*. In the following, we describe the means of such output comparisons, as well as several evaluation metrics.

2.2.2.1. Basic Concepts

For the scope of the works presented in this dissertation, model predictions as well as ground truths can be regarded as labels. Accordingly, comparisons in an evaluation are between predicted labels and ground truth labels.

Confusion Matrix The most basic elements in a comparison between predicted labels and actual (ground truth) labels can be arranged in a *confusion matrix*, as shown below for a binary classification with labels Positive (Pos.) and Negative (Neg.).

		Predicted	
		Pos.	Neg.
Actual	Pos.	TP	FN
	Neg.	FP	TN

Based on the True Positives (TP), False Positives (FP), False Negatives (FN) and True Negatives (TN), different metrics can be calculated.

Accuracy expresses the overall ratio of correct predictions in an evaluation, calculated over both the correct positive, and the correct negative predictions.

$$\text{Accuracy (ACC)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Accuracy is informative in evaluations where the labels are balanced, but of limited information value when this is not the case. In such cases, the metrics precision and recall should be regarded.

Precision expresses for the set of positive predictions, what ratio of these was made correctly.

$$\text{Precision (P)} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall expresses for the set of ground truth positives, what ratio of these was correctly predicted positive.

$$\text{Recall (R)} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Achieving high precision and recall with a model is always a trade-off. This can be illustrated by the fact that a model, which always predicts a positive label, will achieve a perfect recall of 1.¹⁹ Such a model would predict all ground truth positives correctly, but in turn predict all ground truth negatives incorrectly.

¹⁹ Always predicting a positive label ensures FN = 0, which means $R = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{TP}}{\text{TP} + 0} = \frac{\text{TP}}{\text{TP}} = 1$.

F-score is a means to quantify precision and recall in a single value. Its general form, F_β , allows putting more weight on either the precision or recall. Setting β to 1 gives the F_1 -score, which is the harmonic mean of precision and recall.

$$F_1\text{-score } (F_1) = \frac{2 \cdot P \cdot R}{P + R}$$

Precision, recall and, F_1 -score are the typical metrics to assess and compare approaches to the information extraction tasks covered in this dissertation. In the following section, task specific considerations regarding model evaluation are discussed.

2.2.2.2. Task Specific Considerations

Here we briefly present considerations regarding the evaluation of models for specific tasks. Detailed discussions can be found in the respective chapters.

Reference Linking is a task that aims to identify for a bibliographic reference, which publication it is referring to. The task is necessary because references in available data sources (PDF, \LaTeX , etc.) are first and foremost just ambiguous character strings, as they do not unanimously follow a single, rigorously defined format. Furthermore, identifying the referenced publication presupposes a given set of distinct publications (a “target set”) in which the correct one is contained and can be identified. In an idealized model world, this target set encompasses everything that was ever referenced. Practically though, the target set is a finite set of publication records acquired from some data source. For the evaluation of reference linking models, this leads to the following consideration.

If a reference cannot be linked to the target set, it stands to question if

- (a) the reference publication *is* part of the target set, but the matching model failed to identify it.
- (b) the reference publication *is not* part of the target set, and the matching model correctly determined that none of the publications in the target set is the correct match.

As a consequence, the recall of a reference linking model (i.e., the ratio of successfully matched references) is influenced both by the model performance *and* the selection of target set. Recall values should therefore be interpreted with comprehensiveness of the target set in mind. Precision values (i.e., the correctness ratio of matched references), on the other hand, can be understood as is without special considerations.

Entity Recognition (ER) is the task of identifying entities mentioned in a text.²⁰ A key consideration regarding the evaluation of ER models is whether to require *exact matches* of entity mention spans (strict), or whether to also consider *partial matches* (loose). For example, if a text mentions the “Directory of Open Access Journals” and a model only identifies “Directory of Open Access”, this can be seen as either incorrect (strict) or correct (loose). Another consideration is whether an evaluation scheme requires the model, in addition to identifying entity mentions, to also determine each entity’s type, which is common practice. To ensure valid ER model comparisons in our work, we therefore make sure to always compare the same type of measurements (e.g. strict+type with strict+type). The metrics typically used for ER evaluations are precision, recall, and F₁-score [72].

Relation Extraction (RE) has the goal of determining the relations between entities that are mentioned in a text. It can be posed as a classification task on entity pairs. That is, for each possible combination of two entities a and b in a document, the task is to determine a relation label for (a, b) , where “no relation” is among the possibilities. A further distinction can be made by defining a “none of the above” (NOTA) label, which means the two entities *are* in a relation, but there is not fitting label in the set of labels defined by the task [23]. Relation extraction, as described above, is evaluated on an *entity level*, while *mentions* of those entities are contained in the text based on which relations get determined. To illustrate this, consider the following text.

“The Directory of Open Access Journals (DOAJ) currently indexes over 20k journals. Lars Bjørnshauge, who founded the DOAJ 20 years ago, is happy about that.”

A relation $e_{lars} \xrightarrow{\text{founded}} e_{dir}$ exists between the entities e_{lars} , mentioned as “Lars Bjørnshauge”, and e_{dir} , mentioned as “The Directory of Open Access Journals (DOAJ)” and later “DOAJ”. The second sentence containing the entity mentions “Lars Bjørnshauge” and “DOAJ” can be regarded as the *relation evidence*. In our work, we model data for RE tasks *with* relation evidence information. The setting of evaluations we perform depends on the specifics of the models compared to. The metrics typically used for RE evaluations are precision, recall, and F₁-score [145].

2.2.2.3. Bibliographic Note

A more general and extensive introduction to above and related concepts, with extensive pointers to further literature, can be found in [16].

²⁰ It is also referred to as “Named Entity Recognition” (NER). The term “Named Entity Recognition and Classification” (NERC) is sometimes used to distinguish it from only identifying mentions without assigning an entity type. Similarly, “Named Entity Recognition and Disambiguation” (NERD) means entities are furthermore disambiguated, which usually is done by linking to a knowledge graph.

3

Corpus

This chapter is based on the following publication.



Tarek Saier and Michael Färber. “unarXive: A Large Scholarly Data Set with Publications’ Full-Text, Annotated In-Text Citations, and Links to Metadata”. In: *Scientometrics* 125.3 (Dec. 2020), pp. 3085–3108. ISSN: 1588-2861. DOI: 10.1007/s11192-020-03382-z

Remark on the connection to previous work:

The publication cited above is a journal article that represents an extension of a previously published workshop paper [7] (see also Table 1.2). For the journal article, the underlying work, writing, and publication were conducted within the doctoral research period. The preceding workshop paper reports on a result of the master’s thesis before the doctoral research period (see also Section 1.6). Later in this chapter, at the end of Section 3.3, further details regarding the nature of the extension are provided.

The work in this chapter addresses the following research task.

- ❖ **RT1:** *Base Methodology* - establish a base methodology for generating a large-scale, high-quality scholarly data set, that is on par with or improving upon existing data sets.

It furthermore makes contributions to the following research task, which is likewise addressed in the next chapter.

- ❖ **RT2:** *Citation Network Completeness* - develop a method to link literature references, that is able to link more references than are linked in existing data sets, while not compromising on link correctness or processing efficiency.

3.1. Overview

In this chapter, we introduce a methodology for creating a large-scale corpus of linked, full-text documents from \LaTeX source files. The resulting corpus, *unarXive*, comprises over one million documents across multiple scientific disciplines, and spans 27 years. It is further on also used as the basis for the research presented in the subsequent chapters.

Along with the corpus creation methodology, extensive analyses of the resulting corpus are presented.

At the end of the chapter, in Section 3.9, we assess the achievement of the research tasks, as well as the contributions made in terms of the overarching research goal of enabling higher-quality scholarly data.

3.2. Introduction

A variety of approaches exist that utilize scientific paper collections to help researchers in their work. Research paper recommender systems, for example, suggest relevant papers to read [26]. Other systems operate on a more fine-grained level within the full-text of papers, such as the textual contexts in which citations appear (“citation contexts”). Based on citation contexts, things like the citation function [184, 183, 141], the citation polarity [67, 14], and the citation importance [187, 35] can be determined. Furthermore, citation contexts are necessary for context-aware citation recommendation [79, 53], as well as for citation-based document summarization tasks [36], such as citation-based automated survey generation [138] and automated related work section generation [39].

The evaluation of approaches developed for all these tasks, as well as the actual applicability and usefulness of developed systems in real-world scenarios, heavily depend on the data that is used. This typically is a collection of papers provided as full-text, or a set of already extracted citation contexts, consisting of, for instance, 1–3 sentences each. Existing data sets, however, are limited in various ways (see Section 3.3 for more details):

- (1) *Size*. Data sets can be comparatively small (fewer than 100,000 documents), which makes them difficult to use for training and testing machine learning approaches.
- (2) *Cleanliness*. Papers’ full-texts or citation contexts are often very noisy due to the conversion from PDF to plain text and due to encoding issues.
- (3) *Global citation annotations*. Links from citations in the text to the structured representations of the cited publications across documents are often not provided.
- (4) *Data set interlinkage*. For citing and cited documents, data sets often do not provide identifiers from widely used bibliographic databases, such as DBLP¹ or the Microsoft Academic Graph (MAG).²
- (5) *Cross-domain coverage*. Often, only documents from a single scientific discipline are included, restricting the coverage of evaluations and applications based on the data set.

¹ See <https://dblp.uni-trier.de/> [last accessed: 2023-11-06].

² See <https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/> [last accessed: 2023-11-06].

Table 3.1.: Overview of the proposed data set.

	citing documents	references		cited documents
		outgoing	incoming	
<i>full data set:</i>	1,043,126	15,954,664	15,954,664	2,746,288
full-text	1,043,126	15,954,664	7,181,576	736,597
linked to MAG	994,351	15,846,351	15,954,664	2,746,288
<i>by discipline:</i>				
physics	662,894	9,300,576	7,827,072	921,852
mathematics	237,422	3,426,117	5,062,033	906,301
computer science	111,694	2,526,656	1,876,401	425,860
other	31,116	701,315	1,189,158	492,275

data: <http://doi.org/10.5281/zenodo.3385851> [last accessed: 2023-11-06]

code: <https://github.com/111Depence/unarXive> [last accessed: 2023-11-06]

To address these limitations, we propose a new scholarly data set, which we call *unarXive*.³ The data set comprises papers’ full-text, a citation network, and additional metadata. It is freely available at <http://doi.org/10.5281/zenodo.3385851> [last accessed: 2023-11-06] and the implementation for creating it at <https://github.com/111Depence/unarXive> [last accessed: 2023-11-06].

Table 3.1 gives an overview of the proposed data set. Note that throughout this chapter, we refer to links between publications on the *document level* as “references” (corresponding to entries in a section “bibliography” or “references” near the end of a document), whereas on the *text level*, we speak of “citations” (indicated by markers within the text associated with a reference). The proposed data set consists of over one million full-text documents (about 269 million sentences) and links to 2.7 million unique publications via 15.9 million unique references and 29.2 million citations. Thus, we argue that it is considerably large, addressing limitation (1). By using publications’ \LaTeX source files and developing a highly accurate transformation method that converts \LaTeX to plain text, we can resolve issue (2). Besides the pure papers’ content, in-text citations are annotated directly in the text via global identifiers, thereby covering aspect (3). As far as possible, (citing and cited) documents are linked to the Microsoft Academic Graph [173] addressing limitation item (4). This enables us to use the arXiv paper content in combination with the metadata in the MAG, which, as of February 2019, contains data on 213 million publications along with metadata about researchers, venues, and fields of study. Our data set also resolves issue (5), as all disciplines covered in arXiv are included. This enables researchers to analyze papers from several disciplines and to compare approaches using scholarly data across disciplines.

Considering the application of our data set, we argue that it not only can be used as a new large data set for evaluating paper-based and citation-based approaches with unlimited

³ The name is derived from the name of the data source, *arXiv*, and the verb *to unarchive*, indicating the extraction of files from an archive.

citation context lengths (since the publications' full-text is available), but also be a basis for novel ways of paper analytics within bibliometrics and scientometrics. For instance, based on the citation contexts and the citing and cited papers' metadata in the MAG, analyses on biases in the writing and citing behavior of researchers—e.g. related to authors' affiliation [163] or documents' language [115, 120]—can be performed. Furthermore, deep learning approaches, which have been widely used in the digital library domain recently [53], require huge amounts of training data. Our data set allows to overcome this hurdle and investigate how far deep learning approaches can lead us. Overall, we argue that with our data set we can significantly bring the state of the art of big scholarly data one step forward.

We make the following contributions in this chapter:

1. We propose a large, interlinked scholarly data set with papers' full-text, annotated in-text citations, and links to rich metadata. We describe its creation process in detail and provide both the data as well as the creation process implementation to the public.
2. We manually evaluate the validity of our reference links on a sample of 300 references, thereby providing insight into our citation network's quality.
3. We calculate statistical key figures and analyze the data set with respect to its contained references and citations.
4. We compare our reference links to those in the MAG, and manually evaluate the validity of links only appearing in either of the data sets. In doing so, we identify a large number of documents where the MAG lacks coverage.
5. We analyze the likelihood with which in-text citations in our data set refer to specific parts of a cited document depending on the discipline of the citing *and* cited document. Such an analysis is only possible with word level precision citation marker positions annotated in full-text *and* metadata on citing as well as cited documents. The analysis therefore can showcase the practicability of our data set.

The remainder of the chapter is structured as follows: After outlining related data sets in Section 3.3, we describe our data set creation method in Section 3.4. This is followed by statistics and key figures in Section 3.5. In Section 3.6, we evaluate the validity and coverage of our reference links. Section 3.7 is dedicated to the analysis of citation flow and citation contexts. We conclude with a summary and an outlook in Section 3.8, and an overarching result assessment in Section 3.9.

3.3. Existing Data Sets

Table 3.2 gives an overview of related data sets. CiteSeerX can be regarded as the most frequently used evaluation data set for citation-based tasks. For our investigation, we use the snapshot of the entire CiteSeerX data set as of October 2013, published by [83].

Table 3.2.: Overview of existing data sets (#Papers=Number of papers; Cit. contexts=Citation contexts; CS=Computer Science, BM=Biomedicine, LS=Life Sciences, CL=Computer Linguistics; extractable* indicates that extraction might be error-prone due to papers only being available in PDF format).

Data set	# Papers	Cit. contexts	Scope	Full text	Reference IDs
CiteSeerX [34]/RefSeer [83]	1.0 M	400 characters	(all)	no	no
PMC OAS ^a	2.3 M	extractable	BM/LS	yes	mixed
Scholarly Dataset 2 [179]	0.1 M	extractable*	CS	yes	no
arXiv CS [60]	0.09 M	1 sentence	CS	yes	DBLP
ACL-ARC [29]	0.01 M	extractable*	CS/CL	yes	no
ACL-AAN [159]	0.02 M	extractable*	CS/CL	yes	no

^a See <https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/> [last accessed: 2023-11-06].

This data set consists of 1,017,457 papers, together with 10,760,318 automatically extracted citation contexts. This data set has the following drawbacks [164, 60]: The provided meta-information about cited publications is often not accurate. Citing and cited documents are not interlinked to other data sets. Moreover, the citation contexts can contain noise from non-ASCII characters, formulas, section titles, missed references and/or other “unrelated” references, and do not begin with a complete word.

The PubMed Central Open Access Subset (PMC OAS) is another large data set that has been used for citation-based tasks [70, 51, 65]. Contained publications are already processed and available in the JATS [84] XML format. While the data set overall is comparatively clean, heterogeneous annotation of citations within the text and mixed usage of identifiers of cited documents (PubMed, MEDLINE, DOI, etc.) make it difficult to retrieve high-quality citation interlinkings of documents from the data set⁴ [70].

Beside the aforementioned, there are other collections of scientific publications. Among them are the ACL Anthology corpus [29] and Scholarly Dataset 2 [179]. Note that these data sets only contain the publications themselves, typically in PDF format. Therefore, using such data sets for paper-based or citation-based approaches is troublesome, since one must preprocess the data (i.e., (1) extract the content without introducing too much noise, (2) specify global identifiers for cited papers, and (3) annotate citations with those identifiers). Furthermore, there are data sets for evaluating paper recommendation tasks, such as CiteULike⁵ or Mendeley⁶. These, however, only provide metadata about publications or are not freely available for research purposes.

⁴ To be more precise, the heterogeneity makes the usage of the data set *as is* unfeasible. Resolving references to a single consistent set of identifiers retroactively would be an option, but comparatively challenging in the case of PubMed, because of the frequent usage of special notation in publication titles; see also: https://zenodo.org/records/3598421/files/CITREC_Parser_Documentation.pdf [last accessed: 2023-11-06].

⁵ See <https://web.archive.org/web/20190211033621/http://www.citeulike.org/faq/data.adp> [last accessed: 2023-11-06].

⁶ See <https://data.mendeley.com/> [last accessed: 2023-11-06].

Färber et al. also published a data set with annotated arXiv papers’ content in the past [60]. In comparison, our data set is superior to it in the following regards:

- (1) Our data set is considerably larger (1 M instead of 90 k documents).
- (2) Our data set ensures cleanliness of the papers’ full-text and citation contexts by using a more feature rich conversion pipeline.
- (3) We develop a new method for resolving references to consistent global identifiers. Contrary to the method in [60], we evaluate our method and thereby demonstrate its performance (see Section 3.6.1).
- (4) While [60] link documents solely to DBLP, which covers computer science papers, our data set links documents to the Microsoft Academic Graph, which covers all scientific disciplines and which has been used frequently in the digital library domain in recent years [139].
- (5) While the data set in [60] is restricted to computer science, the new data set covers all domains of arXiv (see Section 3.5 and Figure 3.7).

Lastly, compared to a preliminary version of our data set [7], the data we present here has been improved in several ways. Most notably, while in the initial version, only citing papers were associated with arXiv identifiers and only cited papers had been linked to the MAG, we now provide both types of IDs for both sides. This means, that for nearly all documents, MAG metadata is easily accessible, and full-text is not only available for all citing papers but now also for over a quarter of the cited papers. Moreover, we provide significantly more details and insights into the data set’s creation process (see Section 3.4) and its resulting characteristics (see Sections 3.6 and 3.7).

3.4. Data Set Creation

Scientific publications are usually distributed in formats targeted at *human consumption* (e.g., PDF) or, in cases like arXiv, also as source files the aforementioned (e.g., \LaTeX sources for generating PDFs). Citation-based tasks, such as context-aware citation recommendation, in contrast, require *automated processing* of the publications’ textual contents as well as the documents’ interlinking through in-text citations. The creation of a data set for such tasks therefore encompasses two main steps: extraction of plain text and resolution of references. In the following, we will describe how we approached these two steps using arXiv publications’ \LaTeX sources and the Microsoft Academic Graph.

3.4.1. Used Data Sources

The following two resources are the basis of the data set creation process.

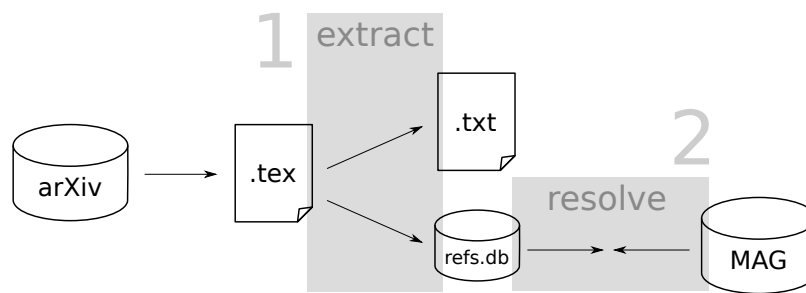


Figure 3.1.: Schematic representation of the data set generation process.

arXiv hosts over 1.5 million documents from August 1991 onward.⁷ They are available not only as PDF, but (in most cases) also as \LaTeX source files. The discipline most prominently represented is physics, followed by mathematics, with computer science seeing a continued increase in percentage of submissions ranking third (see Figure 3.7). The availability of \LaTeX sources makes arXiv documents particularly well suited for extracting high-quality plain text and accurate citation information. So much so, that it has been used to generate ground truths for the evaluation of PDF-to-text conversion tools [24].

Microsoft Academic Graph is a very large, automatically generated data set on 213 million publications, related entities (authors, venues, etc.), and their interconnections through 1.4 billion references.⁸ It has been widely used as a repository of all publications in academia in the fields of bibliometrics and scientometrics [139]. While pre-extracted citing sentences are available, these do not contain annotated citation marker positions. Full text documents are also not available. The size of the MAG makes it a good target for matching reference strings⁹ against it, especially given that arXiv spans several disciplines.

3.4.2. Pipeline Overview

To create the data set, we start out with arXiv sources (see Figure 3.1). From these we generate, per publication, a plain text file with the document’s textual contents and a set of database entries reflecting the document’s reference section. Association between reference strings and in-text citation locations are preserved by placing citation markers in the text. In a second step, we then iterate through all reference strings in the database and match them against paper metadata records in the MAG. This gives us full-text arXiv papers with (word level precision) citation links to MAG paper IDs. As a final step, we enrich the data with MAG IDs on the citing paper side (in addition to the already present arXiv IDs) and arXiv IDs on the cited paper side (in addition to the already present MAG IDs)—this is a straightforward process, because the paper metadata in the MAG includes

⁷ See https://arxiv.org/stats/monthly_submissions [last accessed: 2023-11-06].

⁸ Numbers as of February 2019.

⁹ I.e., the entries in the reference section of a publication. See Listing 3.1 for examples.

Table 3.3.: Comparison of tools for parsing \LaTeX .

Tool	Output	Robust	Usable as is
plastex ^a	DOM	no	yes
TexSoup ^b	document tree	no	yes
opendetex ^c /detex ^d	plain text	no	yes
GrabCite [60]	plain text + resolved ref.	yes	no
\LaTeX XML ^e	XML	yes	yes
Tralics ^f	XML	yes	yes

^a See <https://github.com/tiarno/plastex> [last accessed: 2023-11-06].

^b See <https://github.com/alvinwan/texsoup> [last accessed: 2023-11-06].

^c See <https://github.com/pkubowicz/opendetex> [last accessed: 2023-11-06].

^d See <https://www.freebsd.org/cgi/man.cgi?query=detex> [last accessed: 2023-11-06].

^e See <https://github.com/bruceMiller/LaTeXXML> [last accessed: 2023-11-06].

^f See <https://www-sop.inria.fr/marelle/tralics/> [last accessed: 2023-11-06].

source URLs, meaning papers found on arXiv have an arXiv.org source URL associated with them, such that a mapping from arXiv IDs to MAG IDs can be created.

Listing 3.2 shows how our data set looks like. In the following, we describe the main steps of the data set creation process in more detail.

3.4.3. \LaTeX Parsing

In the following, we will describe the tools considered for parsing \LaTeX , the challenges we faced in general and with regard to arXiv sources in particular, and our resulting approach.

3.4.3.1. Tools

We took several tools for a direct conversion from \LaTeX to plain text or to intermediate formats into consideration and evaluated them. Table 3.3 gives an overview of our results. Half of the tools failed to produce any output for a large amount of arXiv documents we used as test input and were therefore deemed not robust enough. *GrabCite* [60] is able to parse 78.5% of arXiv CS documents but integrates resolving references (see Section 3.4.4) against DBLP into the parsing process and therefore would require significant modification to fit our new system architecture. *LaTeXXML* and *Tralics* are both robust and can be used as \LaTeX conversion tools as is. Based on subsequent tests, we observed that *LaTeXXML* needs

on average 7.7 seconds (3.3 if formula environments are heuristically removed beforehand) to parse an arXiv paper, while *Tralics* needs 0.09. Because the quality of their output seemed comparable, we chose to use *Tralics*.

3.4.3.2. Challenges

Apart from the general difficulty of parsing \LaTeX due to its feature richness and people’s free-spirited use of it, we especially note difficulty in dealing with extra packages not included in documents’ sources.¹⁰ While *Tralics*, for example, is supposed to deal with *natbib* citations,¹¹ normalization of such citations leads to a decrease of citation markers not being able to be matched to an entry in the document’s reference section from 30% to 5% in a sample of 565,613 citations we tested.

3.4.3.3. Resulting Approach

Our \LaTeX parsing solution consists of three steps: flattening, parsing, and output generation. First, we flatten each arXiv document’s sources to a single \LaTeX file using *latexexpand*^{12,13} and normalize citation commands (e.g. `\citep*`, `\citet[see]`, `\citealt`, etc. to `\cite`) to prevent parsing problems later on. In the second step, we then generate an XML representation of the \LaTeX document using *Tralics*. Lastly, we go through the generated XML structure and produce two types of output—(i) an annotated plain text file with the document’s textual contents and (ii) database entries reflecting the document’s reference section. For (i) we replace XML nodes that represent formulas, figures, tables, as well as intra-document references with replacement tokens and turn XML nodes originating from citation markers in the \LaTeX source (i.e., `\cite`) into plain text citation annotation markers. For (ii), each entry in the document’s reference section is assigned a unique identifier, its text is stored in a database, and the identifier put into the corresponding annotation in the plain text (see Listing 3.2).

3.4.4. Reference Resolution

Resolving references to globally consistent identifiers (e.g. detecting that the reference strings (1), (2), and (3) in Listing 3.1 all reference the same document) is a challenging and still unsolved task [144]. Given it is the most distinctive singular part of a publication, we base our reference resolution on the title of the cited work and use other pieces of

¹⁰ The arXiv guidelines specifically suggest the omission of such (see https://arxiv.org/help/submit_tex#wegotem [last accessed: 2023-11-06]).

¹¹ See <https://www-sop.inria.fr/marelle/tralics/packages.html#natbib> [last accessed: 2023-11-06].

¹² See <https://ctan.org/pkg/latexexpand> [last accessed: 2023-11-06].

¹³ We also tested *flatex* (<https://ctan.org/pkg/flatex> [last accessed: 2023-11-06]) and *flap* (<https://github.com/fchauvel/flap> [last accessed: 2023-11-06]) but got the best results with *latexexpand*.

information (e.g., the authors' names) only in secondary steps. In the following, we will describe the challenges we faced, matching arXiv documents' reference strings against MAG paper records, and how we approached the task.

3.4.4.1. Challenges

Reference resolution can be challenging when reference strings contain only minimal amounts of information, when formulas or other special notation is used in titles, or when they refer to non publications (e.g., Listing 3.1, (4)–(6)). Another problem we encountered was noise in the MAG. One such case are the MAG papers with IDs 2167727518 and 2763160969. Both are identically titled “*Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*” and dated to the year 2012. But while the former is cited 17k times and cites 112 papers within the MAG, the latter is neither cited nor cites any other papers.¹⁴ Taking the number of citations into account when matching references, reduced the number of mismatches in this particular case from 2,918 to 0 and improved the overall quality of matches in general.

Listing 3.1: Examples of reference strings.

```
(1) V. N. Senoguz and Q. Shafi, arXiv:hep-ph/0412102
(2) V.N. Senoguz and Q. Shafi, Phys. Rev. D 71 (2005) 043514.
(3) V. N. Senoguz and Q. Shafi, ''Reheat temperature in supersymmetric hybrid inflation models,'' Phys. Rev. D 71, 043514 (2005) [hep-ph/0412102].
(4) V.Sauli, JHEP 02, 001 (2003).
(5) Aaij, Roel, et al. "Search for the  $B^0_{(s)} \rightarrow \eta^{\prime} \phi$  decay" Journal of High Energy Physics 2017.5 (2017): 158.
(6) According to the numerous discussions with my colleagues <removed> and <removed> an experimental verification of our theoretical predictions is feasible.
```

3.4.4.2. Resulting Approach

Our reference resolution procedure can be broken down in two steps: title identification and matching. If contained in the reference string, title identification is performed based on an arXiv ID or DOI (where we retrieve the title from an arXiv metadata dump or via crossref.org¹⁵); otherwise we use Neural ParsCit [155].¹⁶ The identified title is then

¹⁴ The MAG record with ID 2763160969 appears to be a noisy duplicate caused by a web source with easily misinterpretable author information (only a partial list is displayed).

¹⁵ See <https://www.crossref.org/> [last accessed: 2023-11-06].

¹⁶ For title identification we also considered two other state of the art [185] tools, namely CERMINE [186] and GROBID [123]. However, we found CERMINE to be considerably slower than the other tools. And while GROBID showed comparable speed and output quality in preliminary tests, Neural ParsCit's tag based output format was more straightforward to integrate than the faceted TEI format structures that GROBID's reference parser module returns.

matched against the normalized titles of all publications in the MAG. Resulting candidates are considered, if at least one of the author’s names (as given in the MAG) is present in the reference string. If multiple candidates remain, we judge by the citation count given in the MAG—this particularly helps mitigate matches to rouge almost-duplicate entries in the MAG, which often have few to no citations, like paper 2763160969 mentioned in the previous section.

3.4.5. Result format

Listing 3.2 shows some example content from the data set. In addition to the paper plain text files and the references database, we also provide the citation contexts of all successfully resolved references extracted to a CSV file as well as a script to create custom exports.¹⁷ For the provided CSV export, we set the citation context length to 3 sentences—the sentence containing the citation as well as the one before and after—as used by [181] and [83]. Each line in an export CSV has the following columns: cited MAG ID, adjacent cited MAG IDs, citing MAG ID, cited arXiv ID, adjacent cited arXiv IDs, citing arXiv ID, text (see bottom of Listing 3.2). Citations are deemed adjacent, if they are part of a citation group or are at most 5 characters apart (e.g. “[27,42]”, “[27], [42]” or “[27] and [42]”). The IDs of adjacent cited documents are added, because those documents are cited in an almost identical context (i.e., only a few characters to the left or right).

Listing 3.2: Excerpts from (top to bottom) a paper’s plain text, corresponding entries in the references database, entries in the MAG, and extracted citation context CSV.

```

It has over 79 million images stored at the resolution of FORMULA
. Each image is labeled with one of the 75,062 non-abstract nouns
in English, as listed in the Wordnet{{cite:9ad20b7d-87d1-47f5-aeed
-10a1cf89a2e2}}{{cite: 298db7f5-9ebb-4e98-9ecf-0bdda28a42cb}} lexi
cal database.
-----
[uuid]          [citing..] [cited..]   ... [reference_string]
9ad20b7d-87d1  1412.3684  2081580037  ... George A. Miller (1995)
-47f5-aeed-... . WordNet: A Lexical ..
298db7f5-9ebb  1412.3684  2038721957  ... Christiane Fellbaum (19
-4e98-9ecf-...          98), ""WordNet: An El..
-----
[paperid]      [originaltitle]                [publ..] ..
2038721957    WordNet : an electronic lexical database  MIT Press ..
2081580037    WordNet: a lexical database for English  ACM      ..
-----
2131463865|2038721957|2081580037|1412.3684|||It has over 79 millio
n images stored at the resolution of FORMULA . Each image is label
ed with one of the 75,062 non-abstract nouns in English, as listed
in the Wordnet CIT MAINCIT lexical database. It has been noted th
at many of the labels are not reliable CIT .

```

¹⁷ See Python script `extract_contexts.py` bundled with the data set for details.

3.5. Statistics and Key Figures

In this section we present the data set and its creation process in terms of numbers. Furthermore, insight into the distribution of references and citation contexts is given.

3.5.1. Creation Process

We used an arXiv source dump containing all documents up until the end of 2018 (1,492,923 documents). 114,827 of these were only available in PDF format, leaving 1,378,096 sources. Our pipeline output 1,283,584 (93.1%) plain text files, 1,139,790 (82.7%) of which contained citation markers. The number of reference strings identified is 39,694,083, for which 63,633,427 citation markers were placed within the plain text files. This first part of the process took 67 hours to run, unparallelized on an 8 core Intel Core i7-7700 3.60GHz machine with 64 GB of memory.

Of the 39,694,083 reference strings, we were able to match 16,926,159 (42.64%) to MAG paper records. For 31.32% of the reference strings we could neither find an arXiv ID or DOI, nor was Neural ParsCit able to identify a title.¹⁸ For the remaining 26.04% a title was identified, but could not be matched to the MAG. Of the matched 16.9 million items' titles, 52.60% were identified via Neural ParsCit, 28.31% by DOI and 19.09% by arXiv ID. Of the identified DOIs, 32.9% were found as is, while 67.1% were heuristically determined. This was possible because the DOIs of articles in journals of the American Physical Society follow predictable patterns. The matching process took 119 hours, run in 10 parallel processes on a 64 core Intel Xeon Gold 6130 2.10GHz machine with 500 GB of memory.

Comparing the performance of our approach using all papers (1991–2018) to using only the papers from 2018 (i.e., recent content), we note that the percentage of successfully extracted plain texts goes up from 93.1 to 95.9% (82.7 to 87.8% only counting plain text files containing citation markers) and the percentage of successfully resolved references increases from 42.64 to 59.39%. A possible explanation for the latter would be, that there is more and higher quality metadata coverage (MAG, crossref.org, etc.) of more recent publications.

¹⁸ To assess whether or not the large percentage of reference strings without identified title is due to Neural ParsCit missing a lot of them, we manually check its output for a random sample of 100 papers (4027 reference strings). We find that 99% of cases with no title identified actually do not contain a title—like for example items (1), (2) and (4) in Listing 3.1. These kinds of references seem to be most common in physics papers. The 1% where a title was missed were largely references to non-English titles and books. We therefore conclude that the observed numbers largely reflect the actual state of reference strings rather than problems with the approach taken.

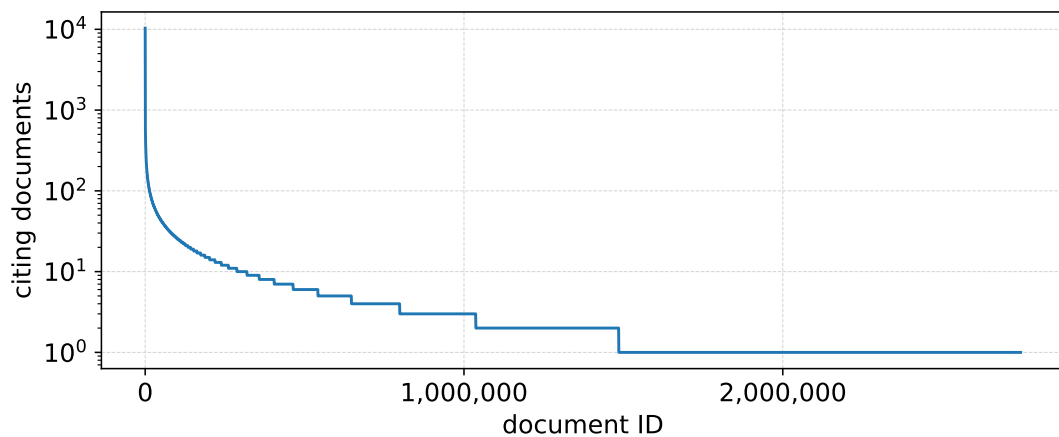


Figure 3.2.: Number of citing documents per cited document.

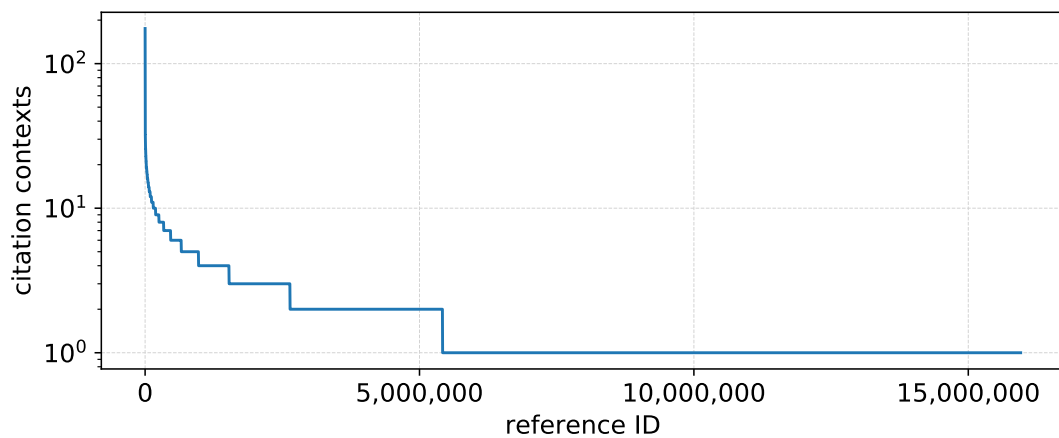


Figure 3.3.: Number of citation contexts per reference.

3.5.2. Resulting Data Set

Our data set consists of *2,746,288 cited papers*, *1,043,126 citing papers*, *15,954,664 references* and *29,203,190 citation contexts*.¹⁹

Figure 3.2 shows the number of citing documents for all cited documents. There is one cited document with over 10,000 citing documents, another 8 with more than 5,000 and another 14 with more than 3,000. 1,485,074 (54.07%) of the cited documents are cited at least two times, 646,509 (23.54%) at least five times. The mean number of citing documents per cited document is 5.81 (SD 28.51). Figure 3.3 shows the number of citation contexts per entry in a document's reference section. 10,537,235 (66.04%) entries have only one citation context, the maximum is 278, the mean 1.83 (SD 2.00).

¹⁹ References that were successfully matched to a MAG record but have no associated citation markers (due to parsing errors; see Section 3.4.3.2) are not counted here.

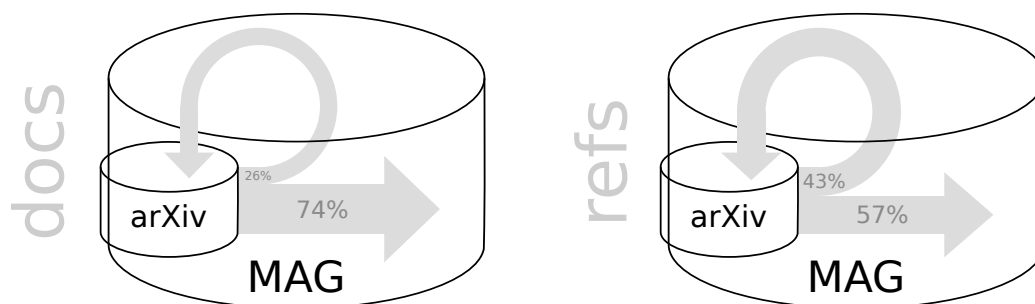


Figure 3.4.: Visualization of the citation flow in terms of documents and references from arXiv to the MAG.

Because not all documents referenced by arXiv papers are hosted on arXiv itself, we additionally visualize the citation flow with respect to the MAG in Figure 3.4. 95% of our citing documents are contained in the MAG. Of the cited documents, 26% are contained in arXiv and therefore included as full-text, while 74% are only included as MAG IDs. On the level of references, this distribution shifts to 43/57. The high percentages of citation links contained within the data set can be explained due to the fact, that in physics and mathematics—which make up a large part of the data set—it is common to self-archive papers on arXiv.

3.6. Evaluation of Citation Data Validity and Coverage

3.6.1. Citation Data Validity

To evaluate the validity of our reference resolution results, we take a random sample of 300 matched reference strings and manually check for each of them, if the correct record in the MAG was identified. This is done by viewing the reference string next to the matched MAG record and verifying, if the former actually refers to the latter.²⁰ Given the 300 items, we observed 3 errors, giving us an accuracy estimate of 96% at the worst, as shown in Table 3.4. Table 3.5 shows the three incorrectly identified documents. In all three cases the misidentified document’s title is contained in the correct document’s title, and there is a large or complete author overlap between correct and actual match. This shows that authors sometimes title follow-up work very similarly, which leads too hard to distinguish cases.

3.6.2. Citation Data Coverage

For the 95% of our data set, where citing as well as cited document have a MAG ID, we are able to compare our citation data directly to the MAG. The composition of reference section

²⁰ Further details can be found at https://github.com/IllDence/unarXive/tree/legacy_2020/doc/matching_evaluation [last accessed: 2023-11-06].

Table 3.4.: Confidence intervals for a sample size of 300 with 297 positive results as given by Wilson score interval and Jeffreys interval [32].

Confidence level	Method	Lower limit	Upper limit
0.99	Wilson	0.9613	0.9975
	Jeffreys	0.9666	0.9983
0.95	Wilson	0.9710	0.9966
	Jeffreys	0.9736	0.9972

Table 3.5.: Mismatched documents.

#	Document
1	matched “ <i>The Maunder Minimum</i> ” (John A. Eddy; 1976) correct “ <i>The Maunder Minimum: A reappraisal</i> ” (John A. Eddy; 1983)
2	matched “ <i>Support Vector Machines</i> ” (Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani; 2013) correct “ <i>1-norm Support Vector Machines</i> ” (Ji Zhu, Saharon Rosset, Robert Tibshirani, Trevor J. Hastie; 2003)
3	matched “ <i>The Putative Liquid-Liquid Transition is a Liquid-Solid Transition in Atomistic Models of Water</i> ” (David Chandler, David Limmer; 2013) correct “ <i>The putative liquid-liquid transition is a liquid-solid transition in atomistic models of water. II</i> ” (David T. Limmer, David Chandler; 2011)

coverage (i.e., how many of the references are reflected in each of the data sets) of all 994,351 citing documents can be seen in Figure 3.5. Of the combined 26,205,834 reference links, 9,829,797 are contained in both data sets (orange), 5,918,128 are in unarXive only (blue), and 10,457,909 are in the MAG only (green). On the document level we observe, that for 401,046 documents unarXive contains more references than the MAG, and for 545,048 it is the other way around. The striking difference between reference and document level²¹ suggests, that the MAG has better coverage of large reference sections. This is supported by the fact that citing papers, where the MAG contains more references, cite on average 34.28 documents, while the same average for citing papers, where unarXive contains more references, is 17.46. Investigating further, in Figure 3.6 we look at the number of citing documents in terms of reference section size (x -axis) and *exclusive coverage in unarXive and MAG*²² (y -axis). As we can see (and as the almost exclusively blue area on the right

²¹ While the number of reference links exclusive to the MAG is about twice as high as the number of reference links exclusive to unarXive, the number of documents for which either of the data sets has better coverage is on a comparable level.

²² Calculated as $\frac{\# \text{citations only in unarXive} - \# \text{citations only in MAG}}{\# \text{citations in both} + \# \text{citations only in unarXive} + \# \text{citations only in MAG}}$.

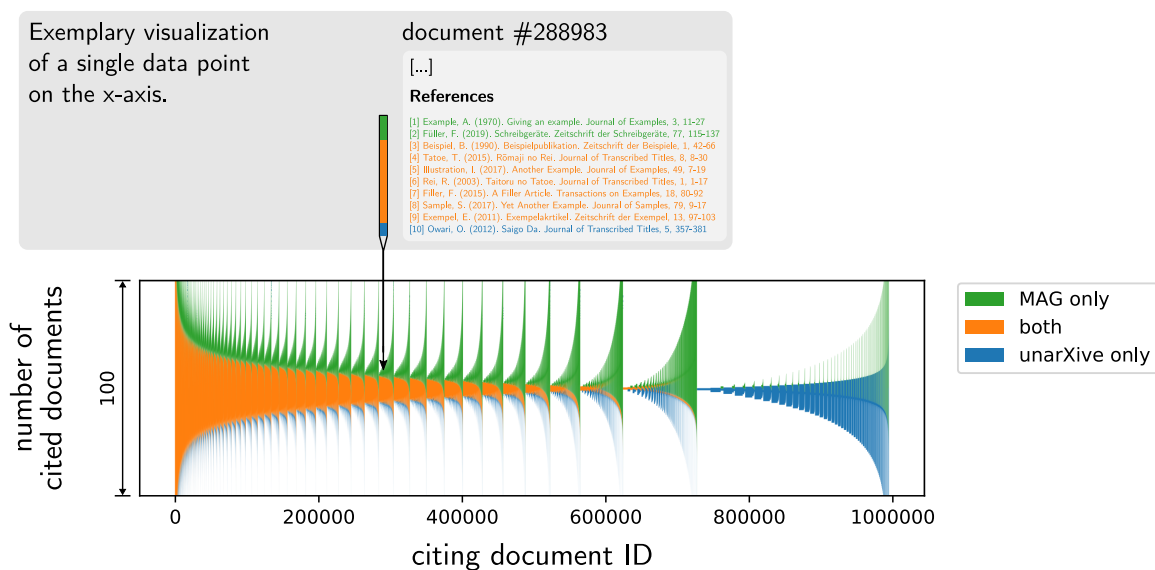


Figure 3.5.: Composition of reference section coverage for all citing documents (cut off at 100 cited documents).

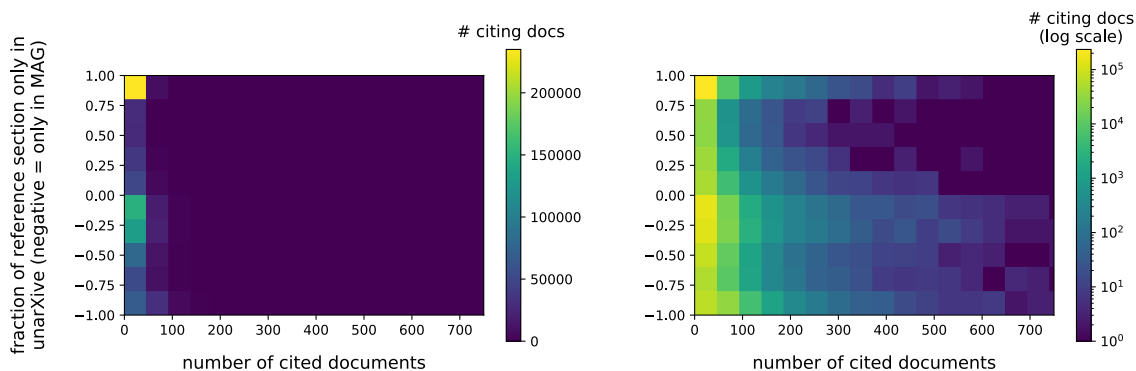


Figure 3.6.: Distribution of citing documents in terms of reference section size and their coverage in unarXive and MAG (cut off at 750 cited documents).

hand side of Figure 3.5 suggests), there is a large number of papers, citing ≤ 50 documents, where $\geq 80\%$ of the reference section are only contained in unarXive. Put differently, there is a large portion of documents, where the reference section is covered to some degree by unarXive, but has close to no coverage in the MAG. The number of citing documents, where the MAG contains 0 references whereas unarXive has ≥ 1 , is 215,291—these have an average of 15.1 references in unarXive.²³ The number of citing documents (within the 994,351 at hand), where unarXive contains 0 references whereas the MAG has ≥ 1 , is 0.

²³ Manually looking into a sample of 100 of these documents, we find the most salient commonality to be irregularities w.r.t. to the reference section headline. 58 of the papers (55 physics, 2 quantitative biology, 1 CS) have no reference section headline, 2 have a double reference section headline and further 2 have the headline directly followed by a page break. The reason for the large number of MAG documents with no references might therefore be, that the PDF parser used can not yet deal with such cases.

Needless to say, additional references are only of value if they are valid. From both the citation links only found in unarXive, as well as those only found in the MAG, we therefore take a sample of 150 citing paper cited paper pairs and manually verify, if the former actually references the latter. This is done by inspecting the citing paper’s PDF and checking the entries in the reference section against the cited paper’s MAG record.²⁴ On the unarXive side, we observe 4 invalid links, all of which are cases similar to those showcased in Table 3.5. On the MAG side, we observe 8 invalid links. Some of them seem to originate from the same challenges as the ones we face, e.g. similarly titled publications by same authors, leading to misidentified *cited* papers. Other error sources are, for instance, an invalid source for a *citing* paper being used and its reference section parsed (e.g. paper ID 1504647293, where one of the PDF sources is the third author’s Ph.D. thesis instead of the described paper). Given that the citation links exclusive to unarXive appear to be half as noisy as those exclusive to the MAG, we argue that the 5,918,128 links only found in unarXive could be useful for citation and paper based tasks using MAG data. This would especially be the case for the field of physics, as it makes up a significant portion of our data set.

3.7. Analysis of Citation Flow and Citation Contexts

Because the documents in unarXive span multiple scientific disciplines, interdisciplinary analyses, such as the calculation of the flow of citations between disciplines, can be performed. Furthermore, the fact that documents are included as full-text and citation markers within the text are linked to their respective cited documents, makes varied and fine-grained study of citation contexts possible. To give further insight into our data set, we therefore conduct several such analyses in the following. Note that, for interdisciplinary investigations, disciplines other than physics, mathematics, and computer science are combined into *other* for space and legibility reasons, as they are only represented by a small number of publications. On the citing documents’ side, these span the fields of economics, electrical engineering and systems science, quantitative biology, quantitative finance, and statistics. Combined on the cited documents’ side are chemistry, biology, engineering, materials science, economics, geology, psychology, medicine, business, geography, sociology, political science, philosophy, environmental science, and art.

3.7.1. Citation Flow

Figure 3.7 depicts the flow of citations by discipline for all 15.9 million matched references. As one would expect, publications in each field are cited the most from within the field itself. Notable is, that the incoming citations in mathematics are the most varied (physics and computer science combined make up 35% of the citations). As citation contexts are

²⁴ Further details can be found at https://github.com/IllDepence/unarXive/tree/legacy_2020/doc/coverage_evaluation [last accessed: 2023-11-06].

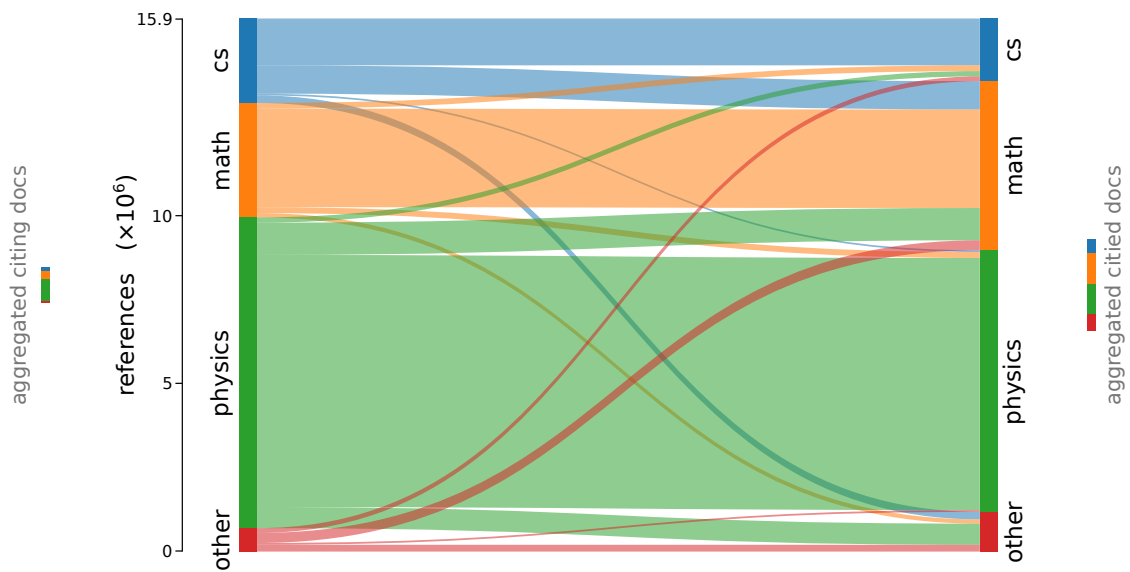


Figure 3.7.: Citation flow by discipline for 15.9 million references (the number of citing and cited documents per discipline are plotted on the sides).

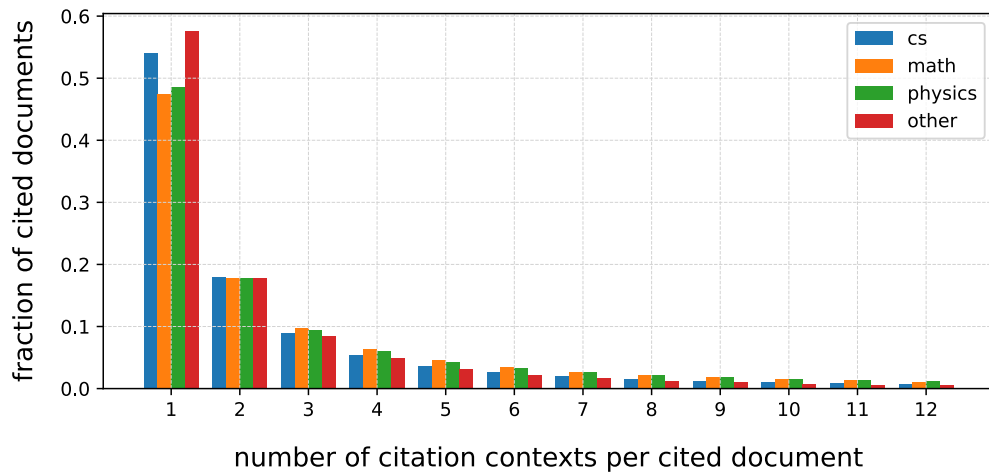


Figure 3.8.: Normalized distribution of the number of citation contexts per cited document.

useful descriptive surrogates of the documents they refer to [55], a composition as varied as mathematics in Figure 3.7 bears the question whether a distinction by discipline could be worth considering, when using citation contexts as descriptions of cited documents. That is, computer scientists and physicists might refer to math papers in a different way than mathematicians do. Borders between disciplines are, however, not necessarily clear-cut, meaning that such a distinction might not be as straight forward as the color coding in Figure 3.7 suggests.

3.7.2. Availability of Citation Contexts

Another aspect that becomes relevant, when using citation contexts to describe cited documents, is the number of citation contexts available per cited publication. Figure 3.8 shows, that the distribution of the number of citation contexts per cited document is similar across disciplines. In each discipline, around half of the cited documents are just mentioned once across all citing documents, 17.5% exactly twice, and so on. The tail of the distribution drops a bit slower for physics and mathematics. The mean values of citation contexts per cited document are 9.5 (SD 50.3) in physics, 7.0 (SD 28.8) in mathematics, 5.1 (SD 31.1) in computer science and 3.5 (SD 11.0) for the combined other fields. This leads to two conclusions. First, it suggests that a representation relying solely on citation contexts may only be viable for a small fraction of publications. Second, the high dispersion in the number of available citation contexts shows that means might not be very informative when it comes to citation counts aggregated over specific sets of documents.

3.7.3. Characteristics of Citation Contexts

For our analysis of the contents of citation contexts, we focus on three aspects: whether or not citations are (1) integral, (2) syntactic and (3) target section specific. These aspects were chosen, because they give particular insights into the citing behavior of researchers, as explained alongside the following definition of terms.

3.7.3.1. “Integral”, “Syntactic” and “Target Section Specific” Citations

We first discuss the terms “*integral*” and “*syntactic*”, which are both established in existing literature. An integral citation is one, where the name of the cited document’s author appears within the citing sentence *and* has a grammatical role [180, 85] (e.g. “Swales [73] has argued that ...”). Similarly, a citation is syntactic, if the *citation marker* has a grammatical role within the citing sentence [194, 15] (e.g. “According to [73] it is ...”). Integral citations are seen as an indication of emphasis towards the cited author (where the opposite direction would be towards the cited work) [180, 85]. Syntactic citations are of interest, when determining how a citation relates to different parts of the citing sentence [194, 15]. Both qualities are relevant when studying the role of citations [59].

Table 3.6 gives a more detailed account of both terms’ use in literature. Note that [108] provide a classification algorithm for integral and non-integral citations that slightly differs from Swales’ original definition depending on the interpretation of a citation marker’s scope, but also gives a clear classification in an edge case where Swales’ definition is unclear. Furthermore note, that the two ways for distinguishing syntactic and non-syntactic citations found in literature are not identical. This is in part because the method given by [15] is kept rather simple. For the intents and purposes of our analysis we follow the definitions of Lamers et al. and Whidby et al. for “*integral*” and “*syntactic*” respectively.

Table 3.6.: Examples of citations and their categorization into integral/non-integral as well as syntactic/non-syntactic (“✓”=yes, “×”=no, “?”=unclear).

Context excerpt (citation marker highlighted)	[180]	[85]	[108]	[194]	[15]
	integral			syntactic	
“Swales (1990) has argued that ...”	✓	✓	✓	×	?
“Swales (1990) has argued that ...”	✓	✓	×	✓	✓
“Swales [73] has argued that ...”	✓	✓	✓	×	×
“Swales has argued that ... [73]”	✓	✓	✓	×	×
“It has been argued (Swales, 1990) that ...”	×	×	×	×	×
“It has been argued [73] that ...”	×	×	×	×	×
“According to (Swales, 1990) it is ...”	?	?	×	✓	✓
“According to [73] it is ...”	×	×	×	✓	✓
“... has been shown (see (Swales, 1990)).”	×	×	×	✓	×

Table 3.7.: Examples of target section specific citations.

Context excerpt (concerns citing document / concerns cited document)
“See [73], Section 3 .”
“This improves Lemma 2 of [73], which is ...”
“Due to this, the proof is now similar to that of Theorem 6.4 from [73].”
“The copolymer version of Theorem 7 was derived in [73], Theorem 3.2 .”
“ Figure 1 is qualitatively similar to Figure 3 in [73].”

As a third aspect for analysis, we define “*target section specific*” citations as those citations, where a specific section within the citation’s target (i.e., the cited document) is referred to. Examples are given in Table 3.7. Target section specific citations are of interest for two reasons. First, in a similar fashion to integral citations, they are a particular form of citing behavior that might be used to infer characteristics of the relationship between citing author and cited document (e.g. a focus on the document rather than authors, or in-depth engagement or familiarity with the cited document’s contents). Second, when using citation contexts as descriptions of cited documents, such as in citation context-based document summarization, target section specific citations might benefit from special handling, as their contexts only describe a (sometimes very narrow) part of the cited document.

In the following we will analyze all three aspects (integral, syntactic, target section specific) with respect to the different scientific disciplines covered by our data set.

Table 3.8.: Per discipline number of citations labeled (1) integral, (2) syntactic, (3) simultaneously integral and syntactic, (4) target section specific (sample size = 300).

Discipline	Integral	Syntactic	Integral+Syntactic	Target Section Specific
Computer Science	23	88	1	5
Mathematics	48	200	13	17
Physics	12	80	2	4
Other	14	113	1	7

3.7.3.2. Manual Analysis of Citation Contexts

For each of the disciplines computer science, mathematics, physics, and other, we take a random sample of 300 citation contexts and manually label them with respect to being integral, syntactic, and target section specific. The result of this analysis is shown in Table 3.8. Each of the assigned labels is most prevalent in mathematics papers, which is furthermore true for the co-occurrence of the labels integral and syntactic. Mathematics is also the only discipline, in which citations are more likely to be syntactic than not. The difference in frequency of integral and syntactic citations might be due to variations in writing culture between the different disciplines. We think that the comparatively high frequency of target section specific citations in mathematics could be due to the fact, that in mathematics intermediate results like corollaries and lemmata are immediately reusable in related work. We further investigate target section specific citations in the following section.

3.7.3.3. Automated Analysis of Target Section Specific Citations

Sentences including a target section specific citation often follow distinct and predictable patterns. For example, a capitalized noun (e.g. “*Corrolary*”, “*Lemma*”, “*Theorem*”) is followed by a number and a preposition (e.g. “*in*”, “*of*”), and then followed by the citation marker (e.g. “*Corrolary 3 in [73]*”). Another pattern is the citation marker followed by a capitalized noun and a number (e.g. “[73] *Lemma 7*”). This lexical regularity allows us to identify target section specific citations in an automated fashion. Specifically, we search the entirety of our 29 M citation contexts for word sequences, that match either of the part of speech tag patterns NNP CD IN <citation marker> and <citation marker> NNP CD. Doing this, we find 365,299 matches (1.25% of all contexts). This is less than the 2.31% one would expect due to the manual analysis²⁵ and suggests, that above two patterns are not exhaustive. Nevertheless, we can use the identified contexts to further analyze them with respect to their distribution of disciplines.

²⁵ Because disciplines are not equally represented in the data set, the expected value is not simply the average of values in Table 3.8 ($\frac{5+17+4+7}{4} \times 300^{-1} = 0.0275$), but a weighted average ($5 \times w_{cs} + 17 \times w_{math} + 4 \times w_{phys} + 7 \times w_{other}$) $\times 300^{-1}$, with $\sum w_{\langle \text{discipline} \rangle} = 1$. This gives a value of ≈ 0.0231 .

Table 3.9.: Occurrence of target section specific citations by discipline (pairs annotated as follows, \dagger : Mathematics citing document, \ddagger : Mathematics cited document, $\underline{X \rightarrow X}$: Citing and cited document are from the same discipline).

	Discipline	Count	Normalization factor	Normalized ratio (%)
Citing	Mathematics	298,009	4.66	8.70
	CS	9,123	6.31	0.36
	Physics	30,593	1.72	0.33
Cited	Mathematics	313,651	3.15	6.20
	CS	12,179	8.50	0.65
	Physics	31,087	2.04	0.40
Pairs	$\underline{\text{Math}^\dagger \rightarrow \text{Math}^\ddagger}$	200,859	5.41	6.81
	$\text{Math}^\dagger \rightarrow \text{CS}$	5,134	92.13	2.96
	$\text{Math}^\dagger \rightarrow \text{Phys}$	3,114	89.88	1.75
	$\text{CS} \rightarrow \text{Math}^\ddagger$	3,456	18.82	0.41
	$\text{Phys} \rightarrow \text{Math}^\ddagger$	3,859	16.49	0.40
	$\underline{\text{CS} \rightarrow \text{CS}}$	2,500	11.38	0.18
	$\underline{\text{Phys} \rightarrow \text{Phys}}$	10,374	2.12	0.14
	$\text{CS} \rightarrow \text{Phys}$	50	307.16	0.10
	$\text{Phys} \rightarrow \text{CS}$	137	101.40	0.09

Table 3.9 shows the results of this subsequent analysis. Because our data set does not contain equal numbers of citations from each discipline (see Figure 3.7), we normalize the absolute numbers of pattern occurrences. Rows are then sorted by normalized ratio in decreasing order. Looking at the citing documents (those in which the pattern was found), we see a similar picture to the one in our manual analysis (shown in Table 3.8). Namely, mathematics with the highest count of target section specific citations by far, and a similar count for computer science and physics, where the latter is slightly lower. Counting by the cited documents (the document in which a specific part is being referenced), the differences decrease a little bit, but mathematics still occurs most frequently by far.

An interesting pattern emerges, when taking an even more detailed look and breaking these citations down by the disciplines on *both* sides of the citation relation. We then can observe the following.

- The most determining factor for target section specific citations seems to be, that a mathematician is writing the document.[†] As with integral and syntactic citations, the writing culture of the field might play a role here.
- The second most determining factor then appears to be, that a mathematical paper is being cited.[‡] Mathematics documents might lend themselves to being cited in this way.
- The third most determining factor is an intra-discipline citation (i.e., the citing document is from that same discipline as the cited). This supports the interpretation of

target section specific citations as a sign of familiarity with what is being cited (see Section 3.7.3.1).

Math→Math pairs, where all three of the above factors come into play simultaneously, consequentially show the highest occurrence of target section specific citations by far.

To summarize the results of our analysis of citation flow and citation contexts, we note the following points.

- Publications in mathematics are cited from “outside the field” (e.g. by computer science or physics papers) to a comparatively high degree. Distinguishing citation contexts referring to mathematics publications by discipline might therefore be beneficial in certain applications (e.g. citation-based automated survey generation).
- For most publications, only one or a few citation contexts are available.
- Integral citations appear to be about twice as common in computer science as they are in physics, and again twice as common in mathematics as they are in computer science. Going with Swale’s interpretation of the phenomenon, this would mean the focus put on authors in mathematics is higher than in computer science, and higher in computer science than in physics.
- In mathematics, syntactic citations seem to be more common than non-syntactic citations. This is beneficial for reference scope identification [15] and any sophisticated approaches based on citation contexts (like context-aware citation recommendation), as citation markers in syntactic citations stand in a grammatical relation to their surrounding words.
- We define target section specific citations as those citations, where a specific section within the cited document is referred to. This type of citation is the most common in mathematics (comparing mathematics, computer science and physics). Through a subsequent analysis of 365k target section specific citations, we find that they are more common in intra-discipline citations than in inter-discipline citations. This supports our assumption that they are an indicator for familiarity with the cited document.

Our work regarding the five aspects outlined in the beginning, namely *size*, *cleanliness*, *global citation annotations*, *data set interlinkage*, *cross-domain coverage*, enabled above results. Without sufficient size, our results would be less informative. If our documents contained too much noise, the quality of reference resolution would have deteriorated. Global citation annotations, especially because of their word level precision, make fine-grained lexical analyses of citation contexts like the one in Section 3.7.3.3 possible. Without interlinking our data set to the MAG, available metadata would have been scarce. While we mainly focused on the scientific discipline information in the MAG, there is much more (authors, venues, etc.) that can be worked with in future analyses. Lastly, if our data set would have only covered a single scientific discipline, an analysis of citation flow, as well as interdisciplinary comparisons of citation context criteria would not have been possible.

3.8. Conclusion

Evaluating and applying approaches to research paper-based and citation-based tasks typically requires large, high-quality, citation-annotated, interlinked data sets. In this chapter, we proposed a new data set with over one million papers' full-text, 29.2 million annotated citations, and 29.2 million extracted citation contexts (of three sentences each), ready to be used by researchers and practitioners. We provide the data set and the implementation for creating the data set from arXiv source files online for further usage.

Author Contributions

Tarek Saier: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. Michael Färber: Supervision, Writing – review & editing.

3.9. Result Assessment

The work in this chapter primarily addresses the following research task.

- ✦ **RT1:** *Base Methodology* - establish a base methodology for generating a large-scale, high-quality scholarly data set, that is on par with or improving upon existing data sets.

The presented methodology and resulting corpus *unarXive* are on par or improve upon the identified related work both in terms of scale and quality. Regarding scale, *unarXive* (1 M documents) is among the top three alongside CiteSeerX (1 M) and the PMC OAS (2.3 M). Considering quality, using \LaTeX source files ensures less noisy full-text compared to data sets generated from PDFs, and the established citation network links proved to be highly accurate in a manual evaluation. Accordingly, we deem ✦ **RT1** successfully achieved.

The work presented in this chapter furthermore makes contributions to the following research task.

- ✦ **RT2:** *Citation Network Completeness* - develop a method to link literature references, that is able to link more references than are linked in existing data sets, while not compromising on link correctness or processing efficiency.

The reference linking method developed for *unarXive* is able to link 42.6% of references successfully to the MAG. While no data set with which a direct comparison would be possible existed at the time of publication, the number compares favorably to arXiv CS achieving 39.3%²⁶, and also can be considered an improvement over the PMC OAS with

²⁶ 962,084 out of 2,448,826 references are reported to be successfully linked in the arXiv CS paper [60].

no consistent citation network as well as CiteSeerX with no assessment resented for its citation network. Accordingly, we deem this a significant contribution to **RT2**.

In terms of the overarching research goal of enabling higher-quality scholarly data (see Table 2.4 in Chapter 2), the work presented in this chapter makes the following contributions.

Scholarly Data Quality Contributions - [1]	
Crit.	Contribution
Rel_{CN}	First data set based on all of arXiv with a citation network
Acc_{CN}	> 96% accurate reference matching method (SOTA)
Acc_{SDR}	Low-noise text extraction by using \LaTeX as data source
$\text{Tim}_{\text{C/S}}$	Publications included up until end of the most recent full year
Coy_{CN}	MAG and arXiv IDs included; DOIs linked through MAG records
Cos_{CN}	42.6% reference matching success rate (SOTA)
Cos_{SDR}	Full-text of documents included

Rel_{CN} We provide the first data set based on all of arXiv with a citation network. Previous data sets only cover part of arXiv [60], or do not include a citation network [68]. By covering all of arXiv, the data is of high relevance for use cases focussing on physics, mathematics, or computer science. Because documents submitted to arXiv undergo a moderation process²⁷ in which they are assigned to a topic according to the arXiv taxonomy,²⁸ a fine-granular and reliable determination of relevance to a subject of study is possible. While documents on arXiv are by designation preprints, most of them are self-archived author copies which later appear in peer-reviewed venues—between 75% and 80% averaged over all disciplines, measured on all papers from 2008 to 2017 [119], and at 90.1% in computer science measured on a sample of 18 thousand papers from 2022 [22].

Acc_{CN} Our reference linking method is evaluated at an accuracy of > 96%. By comparison, CiteSeerX [199, 198, 152] provides no assessment of their citation network accuracy, and S2ORC [122] (published shortly after unarXive) only achieves a matching accuracy of 92% on arXiv papers. Our work accordingly achieves state-of-the-art citation network accuracy.

Acc_{SDR} We create document representations not from PDF files but from papers' \LaTeX sources. Text extraction from \LaTeX has been used to generate ground truths for the evaluation of from PDF documents [24]. Accordingly, we argue that our method constitutes an improvement for the accuracy of document representations compared to PDF based approaches.

²⁷ See <https://info.arxiv.org/help/moderation/index.html> [last accessed: 2024-02-03].

²⁸ See https://arxiv.org/category_taxonomy [last accessed: 2024-02-03].

Tim_{C/S} We apply our method for generating scholarly data on all documents on arXiv until end of the most recent full year. Accordingly, the resulting corpus contains more recent documents than data sets released earlier.

Co_{YCN} We provide MAG IDs and arXiv IDs for the documents in our corpus. Furthermore, DOIs are available through the linked MAG paper records. Enabling the use of three different types of unique identifiers makes our data a versatile target for comparing and combining it with other data. Other data sets of comparable size only provide their own identifiers (CiteSeerX) or only feature a heterogeneous set of identifiers (PMC OAS).

Co_{SCN} We are able to successfully link 42.6% of all reference in our data. This makes our citation network more complete than that of comparable existing data sets. Other approaches do not provide an assessment of their citation network completeness (CiteSeerX), or only achieve a lower percentage (arXiv CS achieving 39.3%).

Co_{SDR} For all documents in our data set we provide their full-text content. This means our document representations are more complete than those in metadata sets like the MAG, and on par with other data sets providing full-text such as CiteSeerX and PMC OAS.

4

Reference Coverage and Granularity

This chapter is based on the following publications.



Tarek Saier, Meng Luan, and Michael Färber. “A Blocking-Based Approach to Enhance Large-Scale Reference Linking”. In: *Proceedings of the workshop on understanding literature references in academic full text (ULITE) at JCDL 2022*. June 2022



Tarek Saier, Johan Krause, and Michael Färber. “unarXive 2022: All arXiv Publications Pre-Processed for NLP, Including Structured Full-Text and Citation Network”. In: *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE Computer Society, June 2023, pp. 66–70. DOI: 10.1109/JCDL57899.2023.00020

The work in this chapter addresses the following research task.

- ✦ **RT2:** *Citation Network Completeness* - develop a method to link literature references, that is able to link more references than are linked in existing data sets, while not compromising on link correctness or processing efficiency.

4.1. Overview

In this chapter, we introduce improvements for our corpus creation methodology in two areas. First, we present in Section 4.2 a blocking and matching method applied on the set of references in a corpus. With this, matched references as well as bibliographic couplings can significantly be increased. Second, we present in Section 4.3 an improved conversion method for \LaTeX source files, and with it, an update of the *unarXive* corpus. The improved

corpus creation method enables fine-granularly structured document representations, and achieves a new state-of-the-art reference matching success rate.¹

At the end of the chapter, in Section 4.4, we assess the achievement of the research task, as well as the contributions made in terms of the overarching research goal of enabling higher-quality scholarly data.

4.2. Reference Linking by Inter-Reference Blocking

4.2.1. Introduction

Scholarly data is becoming increasingly important, and with it, its quality and coverage. Connections between publications in the form of literature references are of particular importance, as they are used as a basis for various analyses, decision-making, and applications. Some examples are research output quantification [82], trend detection [37], summarization [55], and recommendation [125, 58].

However, reference linking methods² described in the literature are only able to link around half of the references contained in the original papers to the cited publications [122, 1]. This lack in coverage is especially affecting references to non-English publications [5], which are in general underrepresented in scholarly data [189, 121, 137, 142], along with publications in the humanities [48, 96].

We see the reason for this lack in linked references in two key shortcomings of current methods. First, references are linked using simple string similarity measures that are often relying *only* on publications' title and author information (which is not always contained in references; see Figure 4.1). Second, references are exclusively linked to a target collection of paper records—usually a large metadata set like DBLP³ or OpenAlex,⁴ or a set of IDs like DOIs or PMIDs. This means references to literature which is not contained in the target collection, as well as to non-source items [41], cannot be linked (see “?” markers in Figure 4.2).

Linking references can be seen as a task of entity resolution [44], which is concerned with identifying entities referring to the same object within or between large data sets. Because the task requires a one-to-one comparison between each of the involved entities, it is

¹ After the publication of our initial corpus creation methodology in [1] (see Chapter 3), the now widely used corpus S2ORC [122] adapted our document conversion methodology for the \LaTeX subset of their data set. While they do not achieve a reference matching success rate as high as ours, they made advances regarding the structured document representation. With the work presented in this chapter, we follow and improve upon their example, establishing a fine-granular structured document representation for the *unarXive* corpus.

² In the following, we use “link[ing/ed] references” to refer to connections to cited papers rather than in-text citation markers.

³ See <https://dblp.org/> [last accessed: 2023-11-10].

⁴ See <https://openalex.org/> [last accessed: 2023-11-10].

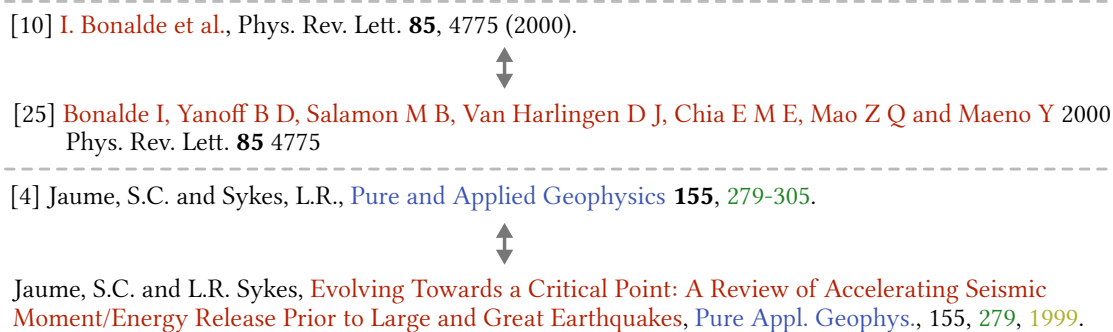


Figure 4.1.: Examples of challenging reference pairs from our evaluation that were successfully matched. **Top:** references from arXiv:cond-mat/0503317 (no title, first author only) and arXiv:cond-mat/0104493 (no title, all authors). **Bottom:** references from arXiv:cond-mat/0104341 (no title, full venue, page range, no year) and arXiv:physics/0504218 (with title, venue abbreviation, start page only, with year).

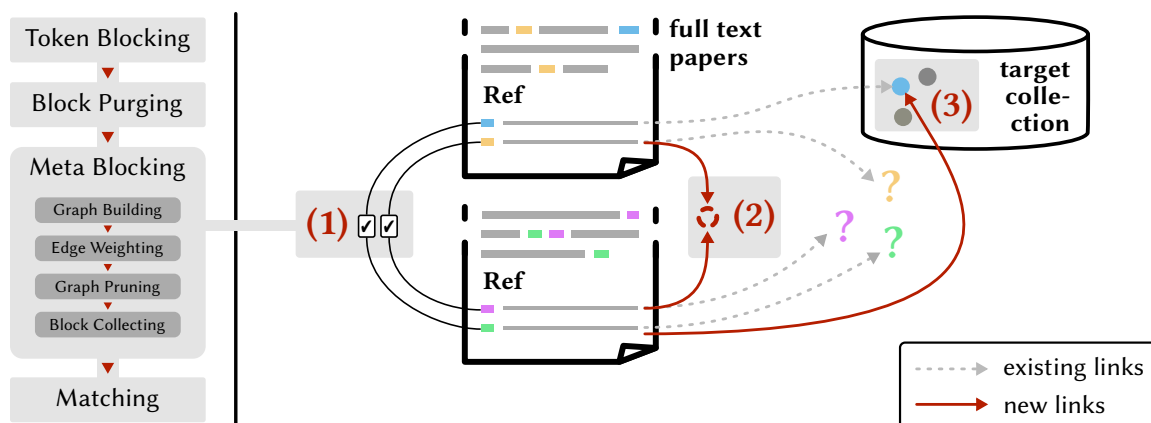


Figure 4.2.: Schematic depiction of the use case. A corpus of full-text papers, where some references are already linked to a target collection (blue), and some are not (orange, pink, green). At (1) we apply our blocking and matching approach to identify all references that point to the same publication. In doing so, we establish new links in the form of (2) bibliographic coupling and (3) links to the target collection.

inherently of quadratic complexity. To make approaches scalable, entities are assigned into groups of likely matching candidates prior to comparison, a technique called blocking [149]. While blocking-based approaches are used in the domain of scholarly data to, for example, identify duplicate paper records [171, 168, 122] (where information such as abstracts are used) and authors [57], they are not utilized among bibliographic references.

We therefore address both of the aforementioned problems of current reference linking approaches, namely, (1) the use of simple matching methods based on title and authors, as well as (2) the reliance on a target collection of paper records, by proposing (1) the use of a blocking and matching process utilizing seven reference fields (title, author, journal, year, etc.) that (2) operates *within* the set of bibliographic references of a corpus, and is thereby *independent* of a target collection of papers (see marker “(1)” in Figure 4.2).

We showcase the feasibility and benefits of our approach, implementing a pre-processing, blocking, and matching pipeline and evaluating it on a corpus containing 300,000 references.

We show that relative to the original data, our approach gives us a *90% increase* in papers linked to the target collection, a *five-fold increase* in bibliographically coupled [31] papers (see marker “(2)” in Figure 4.2), and a *nine-fold increase* in in-text citation markers covered.⁵ The new links are furthermore of high quality (85% F_1 score). This paves the way towards higher quality scholarly data, especially regarding the coverage of so far underrepresented literature and non-source items.

In summary, we make the following contributions.

- We propose a blocking-based approach for matching bibliographic references that is independent of a target collection of paper records.
- We perform a large-scale evaluation showing that our approach results in a manifold increase in high-quality reference links.
- We make our data and code publicly available.⁶

4.2.2. Related Work

Blocking-based approaches have been used in the domain of scholarly data, though to the best of our knowledge not for bibliographic references. We therefore report on (1) exemplary uses of blocking in the scholarly domain for entities other than references, and (2) approaches to linking bibliographic references using methods other than blocking.

Simonini et al. [171] develop BLAST (Blocking with Loosely-Aware Schema Techniques) which adapts Locality-Sensitive Hashing. Among data sets from other domains, they also evaluate their approach for the task of linking 2,600 DBLP paper records to the ACM⁷ and Google Scholar.⁸ Sefid [168] proposes several models to match paper records utilizing the papers’ title, header, and citation information. The models are evaluated in three scenarios matching 1,000 paper records from CiteSeer^x [197] to IEEE, DBLP, and Web of Science. Lastly, Färber et al. [57] detect duplicates among 243 million author records in the Microsoft Academic Knowledge Graph [56] and evaluate their approach using ORCID IDs.

Lo et al. [122] introduced the data set S2ORC, which contains 9.6 million open access papers and has recently seen extensive use in area of scholarly document processing. The authors link references to papers within their data set using a heuristic similarity measure based on n-grams and the Jaccard similarity, which only uses the paper title. Using this method, 26 million out of 50 million references (52%) are successfully linked. The authors report that the low number is “*due to large numbers of papers (mostly in the field of physics) for which the bibliography entries are formatted without paper titles.*” Saier et al. [1]

⁵ With the “coverage” of in-text citation markers we refer to markers associated with linked references, relative to markers belonging to unlinked references.

⁶ See <https://github.com/I11Depence/ulite2022> [last accessed: 2023-11-10].

⁷ See <https://dl.acm.org/> [last accessed: 2023-11-10].

⁸ See <https://scholar.google.com/> [last accessed: 2023-11-10].

introduce unarXive, a data set created from papers' \LaTeX sources containing over 1 million publications. Bibliographic references in the data set are linked to the Microsoft Academic Graph [173, 192]. The linking procedure is based on string similarity of papers' titles and author information. With this procedure 17 million out of 40 million references (42%) are successfully linked. Lastly, CiteSeer^x [199, 197] in another large data set containing paper records. Similar to S2ORC, references are linked to paper records within the data set itself. In the case of CiteSeer^x the linking is performed through a heuristic assignment based on title and author information. We are not aware of information on the percentage of references that are successfully linked in CiteSeer^x.

4.2.3. Approach

Our approach consists of the following three steps: (1) *pre-processing* to convert references into a normalized, structured format, (2) *blocking* to allow us to process large amounts of references, and (3) *matching*. These steps are explained in more detail below.

Pre-processing References as they appear in papers are hard to match for several reasons, such as the variety of citation styles, variants of author names, venue abbreviations, sparsity of information, and typing errors [43] (see Figure 4.1). To mitigate these issues, we pre-process references in three steps: first, we apply GROBID's [123] reference string parsing module,^{9,10} then we expand journal and conference abbreviations, and lastly all strings are lowercased and Unicode normalized. For the abbreviation expansion we use a mapping for 47.6k journal titles provided by JabRef¹¹ and 2.6k conference titles crawled from various web sources. Following [102] we select seven reference fields for the blocking step: title, author, year, volume, journal, booktitle, and pages.

Blocking Following [150], we build our blocking pipeline from components for (1) block building, (2) block cleaning, and (3) comparison cleaning. As shown in Figure 4.2, we use token blocking, block purging, and meta-blocking respectively for each of the steps.

Token blocking is chosen for the block building step because it is schema-agnostic and therefore robust against the varying level of information contained in or missing from bibliographic references. In this step, references are assigned to blocks based on all tokens (i.e., words) contained in the identified and normalized reference fields. As a result, references at this point are associated with multiple blocks, which leads to a high level of redundancy.

Block purging [147] removes oversized blocks based on a comparison cardinality metric, which we determine heuristically and set it to 0.01. Intuitively, the removed blocks originate

⁹ See <https://grobid.readthedocs.io/en/latest/Grobid-service/#apiprocesscitation> [last accessed: 2023-11-10].

¹⁰ GROBID was chosen according to the results of [185].

¹¹ See <https://github.com/JabRef/abbrv.jabref.org> [last accessed: 2023-11-10].

from common tokens, meaning that matched reference strings within them are highly likely to also share smaller blocks. Purging therefore reduces the number of overall comparisons with minimal effect on the final result quality.

Meta-blocking [148], our comparison cleaning step, reduces unnecessary comparisons within blocks by generating a weighted graph of entities (references in our case) based on their shared blocks, removing edges based on a pruning scheme, and lastly creating a new block collection based on the reduced graph. For both the weighting and the pruning of edges several schemes exist. In Section 4.2.4 we describe how we determined the most suitable combination of schemes for our use case. Here, we briefly mention the schemes involved. Available graph weighting schemes include the Common Blocks Scheme (CBS), the Enhanced Common Blocks Scheme (ECBS), the Aggregate Reciprocal Comparisons Scheme (ARCS), and the Jaccard Scheme (JS). For graph pruning, we consider Cardinality Node Pruning (CNP), which relies on cardinality to select the top edges for each node, as well as Weight Edge Pruning (WEP), which removes edges based on their assigned weight.

Matching To determine which references within a block refer to the same publications, we utilize a weighted average of Jaccard similarities across our seven reference fields. Based on [63] as well as preliminary experiments, we set the weights for title, author, journal, booktitle, year, volume, and pages to 8, 6, 5, 5, 3, 3, and 2 respectively, and set the threshold for a match to 0.405.

4.2.4. Evaluation

We use a large corpus of scholarly publications to perform two types of evaluations. (1) A large-scale evaluation utilizing the corpus' existing reference links as ground truth, and (2) a manual evaluation to also assess the correctness of newly created reference links. In the following, we describe the data used, evaluations performed, and results obtained.

Data For our evaluation we use the data set unarXive [1]. We chose this data set over similar data sets such as S2ORC [122], because it not only contains paper's full-text with annotated in-text citation markers, but also a dedicated database of all raw references in plain text. From unarXive we sample the 300,000 most recent references to conduct our evaluation. The 300,000 references originate from 9,917 papers from the disciplines of physics (7,347), mathematics (1,686), computer science (789), and other STEM fields (95). The publications cited through the references cover publication years from 1743 up to 2020. Four examples of references used in the evaluation are shown in Figure 4.1.

Table 4.1.: Performance of five graph weighting and graph pruning scheme combinations for meta-blocking.

Weighting scheme	Pruning scheme	#Comparisons	#Matches	RR ¹ (%)	PC ² (%)	PQ ³ (%)
CBS ⁴	CNP ⁸	39,050	3,053	99.96	54.47	7.82
ECBS ⁵	CNP	39,050	3,201	99.96	57.11	8.20
ARCS ⁶	CNP	39,050	2,890	99.96	51.56	7.40
ARCS	WEP ⁹	24,175	1,285	99.98	22.93	5.32
JS ⁷	WEP	42,919	2,272	99.96	40.54	5.29

Metrics: ¹Reduction Ratio, ²Pair Completeness, ³Pairs Quality

Weighting schemes: ⁴Common Blocks Scheme, ⁵Enhanced Common Blocks Scheme, ⁶Aggregate Reciprocal Comparisons Scheme, ⁷Jaccard Scheme

Pruning schemes: ⁸Cardinality Node Pruning, ⁹Weight Edge Pruning

Large-Scale Evaluation Our large-scale evaluation is performed in two steps. First, we determine the most suitable configuration of graph weighting and pruning scheme for our meta-blocking step, then we apply our pipeline to the evaluation corpus and determine the number of additionally linked entities.

To choose a graph weighting and pruning scheme, we use the 13,976 references in our corpus which are already linked to the target collection as ground truth. Following [148], we select five combinations of schemes to evaluate. The combinations are evaluated using the metrics pair completeness (PC), which expresses the ratio of detected matches with respect to all true matches, pair quality (PQ), which estimates the portion of true matches within all executed comparisons in the block collection, and reduction ratio (RR), which measures the number of unnecessary comparisons that are saved through blocking. Table 4.1 shows the results of our evaluation. We achieve the best results using ECBS weighting and CNP pruning. Accordingly, we apply our pipeline with this configuration on the full evaluation corpus of 300k references, where our approach performs 496,051 comparisons after blocking and identifies 71,826 matches.

As shown earlier in Figure 4.2, we can use the matches identified by our pipeline to create two types of new links. First, new links to the target collection, and second, links between references created through bibliographic coupling. New links to the target collection are established whenever a reference with no existing link is matched to a reference with an existing link (see marker “(3)” in Figure 4.2). In cases where neither of the references in a match have an existing link, we create a bibliographic coupling (see marker “(2)” in Figure 4.2). In Table 4.2 we show on the level of papers, references, and in-text citations how many links were already given in our corpus and how many new links we are able to establish. Regarding links to the target collection, we are able to link *1,443 new papers* (90.75% increase) through *2,442 references* (17.47% increase), which are connected to *7,824 in-text citation markers* (33.00% increase). As for bibliographic coupling, we connect *8,895 papers* through *53,940 references* connected to *219,630 in-text citation markers*. Comparing the number of given links to the combined number of new links, we see a 90% increase in papers linked to the target collection, a five-fold increase in bibliographically coupled papers, and a nine-fold increase in in-text citation markers covered.

Table 4.2.: Number of linked papers, references, and in-text citations given in the original corpus and newly created through the application of our approach.

Linked to target collection			
	#Papers	#Referencecs	#In-text Citations
Given	1,590	13,975	23,707
New	1,443	2,442	7,824
Linked through bibliographic coupling			
	#Papers	#Referencecs	#In-text Citations
Given	-	-	-
New	8,895	53,940	219,630
Combined (linked in either way) ¹			
	#Papers	#Referencecs	#In-text Citations
Given	1,590	13,975	23,707
New	8,931	55,197	227,454

¹ Note that the combined entity counts are not simply the sum of the numbers above, because a single entity can be linked in both ways.

Manual Evaluation To assess the quality of our newly linked references, we take a random sample of 500 reference comparisons from the matching procedure and manually verify if our approach correctly labeled each pair as a match or non-match. This is done by inspecting both original reference strings (prior to pre-processing) and determining whether they refer to the same publication or not. Because in some disciplines such as physics it is common to see references without a title given, this process involves looking up and verifying publications' details online.¹² Examples of two reference pairs are shown in Figure 4.1. Comparing our predicted matches with the manually established ground truth, we measure a precision of 93.20% and a recall of 79.34%. Accordingly the F_1 score is 85.71%. This shows us that our newly established links are of good quality, suggesting our approach facilitates the creation of more accurate scholarly data and, accordingly, higher quality analyses and downstream applications based scholarly data sets.

4.2.5. Discussion and Future Work

To improve the quality of reference linking in large scholarly data sets, we proposed a blocking-based reference linking approach that is independent of a target collection of paper records. In a large-scale evaluation, we first determined the most suitable meta-blocking scheme for our particular application case. Subsequently applying our approach to a corpus of 300,000 references, we saw a manifold increase in linked papers, references,

¹² For further details see https://github.com/IlIDepence/ulite2022/tree/master/5_manual_evaluation [last accessed: 2023-11-10].

and in-text citation markers. The newly established links are of high precision and have a high recall, which we confirmed through a manual evaluation on a sample of our results. This demonstrates the benefits and quality of our approach.

Key limitations of the work presented are (1) the size and discipline coverage of the evaluation corpus, (2) the usage of a comparatively basic blocking technique, and (3) the lack of a thorough evaluation of time performance.

In the future we want to address these points by expanding our work through using more advanced blocking methods such as progressive blocking [172, 64], using larger evaluation corpora such as the whole unarXive data set, including data from more diverse disciplines such as the humanities, and evaluating the time performance of our approach. Because references in our evaluation corpus are linked to in-text citation markers, we furthermore plan to explore application scenarios utilizing the paper full-text.

Author Contributions

Tarek Saier: Conceptualization, Data curation (support), Formal analysis, Investigation (support), Methodology (support), Software (final evaluation), Visualization, Supervision, Writing – original draft (lead), Writing – review & editing. Meng Luan: Data curation, Formal analysis, Investigation, Methodology, Software, Writing – original draft (support). Michael Färber: Supervision, Writing – review & editing.

4.3. Reference & Text Granularity - A Corpus Update

4.3.1. Introduction

Large data sets derived from the full-texts of academic publications are of ever-increasing importance. Beyond large-scale metadata, which is the basis for bibliometric analyses, research output quantification [82], and various applications such as trend detection [37], data sets reflecting the *full-text* content of papers have recently enabled more sophisticated analyses and applications, such as scientific document summarization [126], claim verification [191], and knowledge graph generation [124].

Key aspects of such data sets are (1) basic measures such as quality, size, and temporal as well as disciplinary coverage, (2) their citation network, and (3) handling of non-textual content. (1) Quality is affected by the source material (e.g. PDF or \LaTeX) and parsing method. (2) The citation network is important to allow for bibliometric analyses. (3) Non-textual content such as tables, figures, and mathematical notation often contain important information.

Across these key aspects, we see significant shortcomings in currently available data sets, as shown in Table 4.3. For example, (1) limited size (SciXGen), (2) omission of a citation

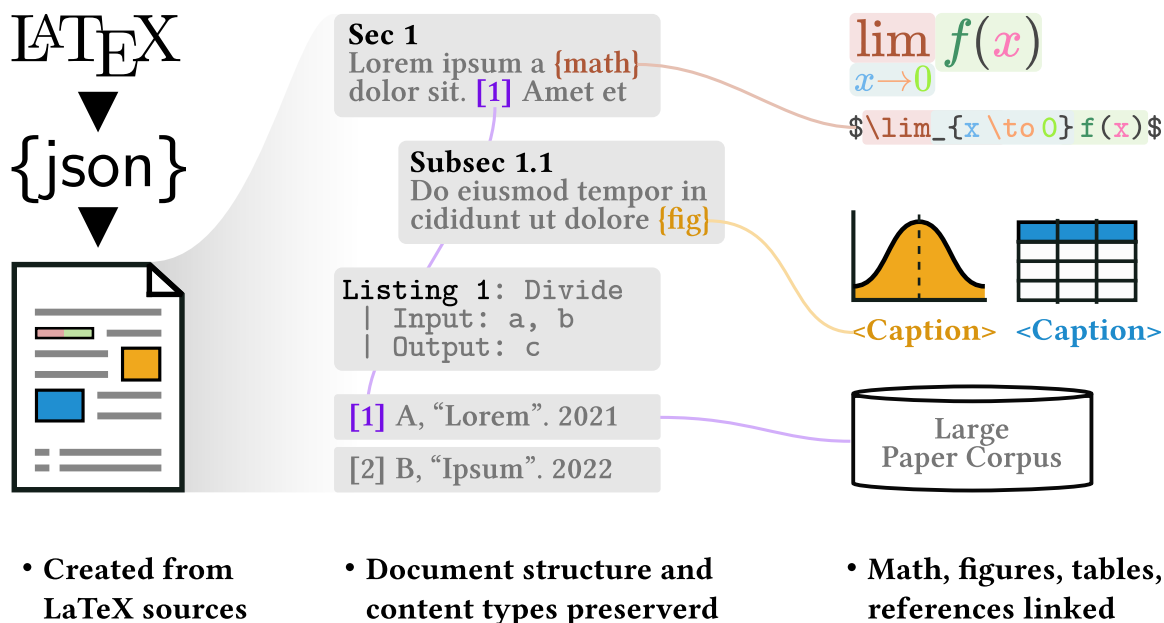


Figure 4.3.: Schematic of our data set. (Created from arXiv.org \LaTeX sources, our data set preserves document structure (sections, subsections, ...) and content types (paragraphs, listings, ...). In-text positions of mathematical notation, figures, tables and citation markers are linked to \LaTeX math content, figure/table captions, and bibliographic references respectively. Bibliographical references are linked to the large paper corpus OpenAlex.

network (arXMLiv), and (3) no or limited handling of mathematical notation (S2ORC, unarXive 2020).

To address these issues, we propose a new version of the data set unarXive, which comprises 1.9 M publications across several disciplines, includes a more complete citation network than its predecessors, and retains structured mathematical notation as well as table and figure captions (see Figure 4.3). Apart from the data set itself, we furthermore provide ready-to-use training and test data for two NLP tasks. Overall, we make the following contributions.

- We provide a 1.9 M document scholarly data set, containing structured full-text, annotated in-text citations, linked table and figure captions, structured mathematical notation, and a high-quality citation network.
- We provide ready-to-use training/test data for the development and evaluation of approaches to two NLP tasks, namely citation recommendation and IMRaD classification.
- We distribute our data in accordance to the FAIR principles [195] and share our source code freely available under a permissive license.

Table 4.3.: Comparison of large data sets derived from paper full-texts ([†]Citation network completeness is reported in two ways. “general[†]”: the whole data set; not directly comparable. “compare[†]”: for arXiv.org data from 1991–2020; directly comparable. [‡]References in the PMC OAS are partially linked to a mixed set of IDs (PubMed, MEDLINE, DOI) [70]. Therefore there is no single, comprehensive number for its completeness.

Data Set	Source		Citation Network [†]			# Docs	Disciplines	Purpose
	Data	Format	general	compare	Structured			
CORE [154]	multiple	PDF	0%	-	×	>100 M	various	general NLP
S2ORC (PDF) [122]	multiple	PDF	69.4%	-	×	12 M	various	general NLP
unarXive 2020 [1]	arXiv.org	Ⓔ $\mathbb{T}\mathbb{E}\mathbb{X}$	42.6%	42.6%	×	1.2 M	phys., maths, CS	general NLP
S2ORC (Ⓔ $\mathbb{T}\mathbb{E}\mathbb{X}$) [122]	arXiv.org	Ⓔ $\mathbb{T}\mathbb{E}\mathbb{X}$	31.1%	31.1%	✓	1.5 M	phys., maths, CS	general NLP
arXMLiv [68]	arXiv.org	Ⓔ $\mathbb{T}\mathbb{E}\mathbb{X}$	0%	0%	✓	1.6 M	phys., maths, CS	maths linguistics
SciXGen [38]	arXiv.org	Ⓔ $\mathbb{T}\mathbb{E}\mathbb{X}$	41.6%	-	✓	205 k	CS	text generation
PMC OAS ¹⁴	PubMed	XML	mixed [‡]	-	✓	3.3 M	biomedical	not NLP specific
unarXive 2022 (ours)	arXiv.org	Ⓔ $\mathbb{T}\mathbb{E}\mathbb{X}$	44.4%	44.4%	✓	1.9 M	phys., maths, CS	general NLP

4.3.2. Related Work

In Table 4.3 we give an overview of related work. Excluded are data sets that are either just sets of PDFs, or only contain metadata.

CORE [154], while being very large, does not contain a citation network, nor is document structure preserved. S2ORC (PDF) [122] is second in size and, while not directly comparable due to different publications covered, has the most complete citation network. However, mathematical notation is only partially preserved as plain text. unarXive 2020 [1] has the second-highest citation network completeness in direct comparison, but lacks structured content.

The bottom part of the table are data sets with both document structure preserved and structured mathematical notation. S2ORC (\LaTeX) [122] is a discontinued¹³ subset of S2ORC and has a limited citation network, arXMLiv [68] offers the highest level of structure but no citation network, and SciXGen [38] is limited in size. The PMC OAS¹⁴ is comparable to unarXive 2022 in size and structure, but has a partial and mixed citation network.

Overall, unarXive 2022 has the most complete citation network as far as direct comparison is possible, preserves document structure as well as structured mathematical notation, and is the largest data set covering physics, mathematics and computer science.

4.3.3. Approach

We base our data set creation approach in part on S2ORC (\LaTeX) and in part on unarXive 2020. This is motivated as follows.

As shown in Table 4.3, the majority of related data sets is based on paper’s \LaTeX sources—which is less noise-prone than parsing PDFs [24]. Among these, S2ORC (\LaTeX) provides well-structured full-text content usable for a wide variety of applications (see Section 4.3.4.2), while arXMLiv and SciXGen are optimized for special purposes. We therefore base our structured document representation on S2ORC (\LaTeX). Regarding the citation network, however, unarXive 2020 achieves the most high-quality results in direct comparison among existing data sets. We therefore base our citation network creation on unarXive 2020.

Regarding both S2ORC (\LaTeX) and unarXive 2020, we do not just copy, but also improve upon the existing work. To furthermore provide an up-to-date data set, we use as source data all papers on arXiv.org up until the end of 2022.

¹³ Last release including the \LaTeX subset is 2019-09-28, see <https://github.com/allenai/s2orc> [last accessed: 2023-02-12].

¹⁴ See <https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/> [last accessed: 2023-11-06].

Conceptually, our overall data set creation process can be broken down into two major steps, namely document parsing and reference linking. In the following these are described in more detail.

4.3.3.1. Document Parsing

To convert the \LaTeX source of a paper into a format that is well suited for NLP applications and analyses, we follow S2ORC (\LaTeX) and unarXive 2020 and perform the following three steps. First, we flatten the paper’s \LaTeX source into a single `.tex` document using `latexexpand`.¹⁵ Next, we use the tool `Tralics`¹⁶ to convert the \LaTeX source into XML. In the last step, we create an easy to handle JSON structure from the XML.

We adapt and extend the JSON structure of S2ORC as shown in Table 4.4. Adding paper metadata facilitates easier analyses (e.g. for specific or across disciplines). Including information on section numbers and types reflects the document structure more closely (e.g. the nesting structure is not lost). Retaining URLs from embedded links helps with reference linking (see Section 4.3.3.2).

We mark the position of citation markers, tables, figures, and mathematical notation within the running text, and link citations markers to their references, tables and figures to their captions (i.e., textual surrogates of their content), and mathematical notation to its original \LaTeX content.

4.3.3.2. Reference Linking

To add a citation network to the data set, bibliographical references—which at this point are just raw strings of text—need to be associated with the cited documents they are referencing. We follow the methodology of unarXive 2020 and link references to a large corpus of publication metadata. To do this, references are first parsed to determine the contained information (title, authors, year, venue, etc.), which is then matched against the paper records in the large metadata corpus. For these two steps, we make the following changes and improvements to the unarXive 2020 approach.

Parsing unarXive 2020 utilizes the tool `Neural Parscit` [155] for reference parsing and furthermore uses a heuristic procedure to determine identifiers such as DOIs or arXiv IDs found within reference string. We use `GROBID` [123], a more commonly used and actively developed tool. Additionally, we extend the identifier determination heuristics to be more robust and versatile by refining matching patterns and extending them to more citation styles.

¹⁵ See <https://ctan.org/pkg/latexexpand> [last accessed: 2023-11-10].

¹⁶ See <https://www-sop.inria.fr/marelle/tralics/> [last accessed: 2023-11-10].

Table 4.4.: Extension of S2ORC format.

Entity	S2ORC data	Added data
Paper	<ul style="list-style-type: none"> • ID • abstract • full-text (list of paragraphs) • bibliographic references 	<ul style="list-style-type: none"> • Metadata (title, list of authors, discipline, license, version history)
Paragraph	<ul style="list-style-type: none"> • Section title • text 	<ul style="list-style-type: none"> • Section number • Section type (e.g. <i>section</i>, <i>subsection</i>) • Content type (e.g. <i>paragraph</i>, <i>listing</i>, <i>proof</i>)
Bibliographic reference	<ul style="list-style-type: none"> • Parsed reference • ID of cited document 	<ul style="list-style-type: none"> • Raw reference string • List of contained arXiv IDs • List of embedded links (i.e., URLs of clickable links not rendered as text when viewing the document)

Matching unarXive 2020 matches references to paper records in the Microsoft Academic Graph (MAG) [173], which is no longer publicly available. Instead of the MAG, we use OpenAlex [156], the MAG’s open successor provided by the nonprofit organization OurResearch.¹⁷ Choosing OpenAlex allows us to also match references to recent papers, which would not be contained in legacy versions of the MAG. Additionally, the fact that OpenAlex paper records contain a variety of identifiers (e.g. DOI and PubMed ID) facilitates combined and comparative analyses of our data with others. Furthermore, OpenAlex has been deemed better suited for bibliographic analyses than the MAG [166].

4.3.4. Results

In the following, we first present key statistics of our proposed data set. Following that, we explain how the data set can be used for analyses as well as the development of NLP applications, and introduce training/test data for two NLP tasks. Lastly, we describe how the data set is distributed to facilitate easy adoption by the community of researchers and practitioners.

¹⁷ See <https://ourresearch.org/> [last accessed: 2023-11-10].

4.3.4.1. Data Set

Our data set comprises *1,881,346 papers*, which contain a combined *182,586,547 paragraphs*, *63,367,836 references* and *133,744,613 in-text citation markers*. The distribution across disciplines is 57% physics, 20% mathematics, 17% computer science, and a combined 5% for others. We are able to link 28,135,565 references (44.4%) and 64,547,944 (48.3%) in-text citation markers to OpenAlex. As shown in Table 4.3, this makes our citation network more complete than that of existing data sets.

In Listing 4.1 we show an excerpt of our document representation for one paper, showcasing the extracted plain text and structured content.

```

/* ----- example paper (arXiv:2105.05862) ----- */
{ "paper_id": "2105.05862",
  "metadata": { ... },
  "abstract": { ... },
  "body_text": [ ... ],
  "ref_entries": { ... },
  "bib_entries": { ... } }
/* ----- one of the sections in body_text ----- */
{ "section": "Memory wave form",
  "sec_number": "2.1",
  "sec_type": "subsection",
  "content_type": "paragraph",
  "text": "The gauge choice leading us to this solution does not fix
          completely all the gauge freedom and an additional constraint
          should be imposed to leave only the physical degrees of freedom.
          This is done by projecting the source tensor  $S_{\mu\nu}$  into its
          transverse-traceless (TT) components (see for example \[80\]).
          Doing this and without loss of generality, we will use the following
          very well known ansatz for the source term proposed in \[9\]
          "
  "ref_entries": { "entry": { "text": "R. Epstein, The Generation of Gravitational Radiation by Escaping Supernova Neutrinos, Astrophys. J. 223 (1978) 1037.",
                              "start": 87,
                              "end": 117 } }
  "bib_entries": { "entry": { "text": "R. Epstein, The Generation of Gravitational Radiation by Escaping Supernova Neutrinos, Astrophys. J. 223 (1978) 1037.",
                              "start": 87,
                              "end": 117 } }
  "contained_links": [
    { "url": "https://doi.org/10.1086/156337",
      "text": "Astrophys. J. 223 (1978) 1037.",
      "start": 87,
      "end": 117 }
  ],
  "ids": { ... } }

```

Listing 4.1: Data example.

In Figure 4.4 we show the number of papers across all disciplines over all years covered. We can see that yearly arXiv.org submissions in computer science are likely to surpass those in physics in 2023. As a simple showcase of the use of structured full-text content, we show in Figure 4.5 how the average number of bibliographic references per paragraph developed over time for the three major disciplines represented in the data set. Dividing by paragraphs is done to account for variation in paper length. We can see that the density of references is increasing more rapidly in physics and computer science, than it is in mathematics.

4.3.4.2. Applications

As is evident by the past use of our data set's predecessors unarXive 2020 and S2ORC, large-scale scholarly data sets created with NLP research in mind have broad applicability. Example uses are analyses of citation behavior across languages [5] or disciplines [188]

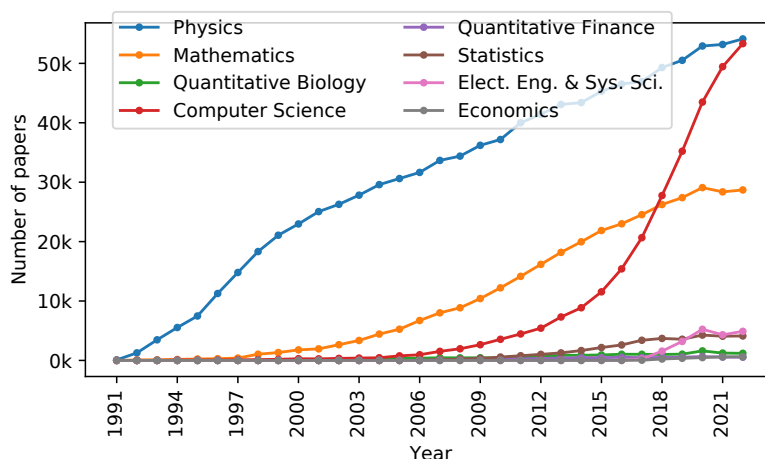


Figure 4.4.: Number of papers per year.

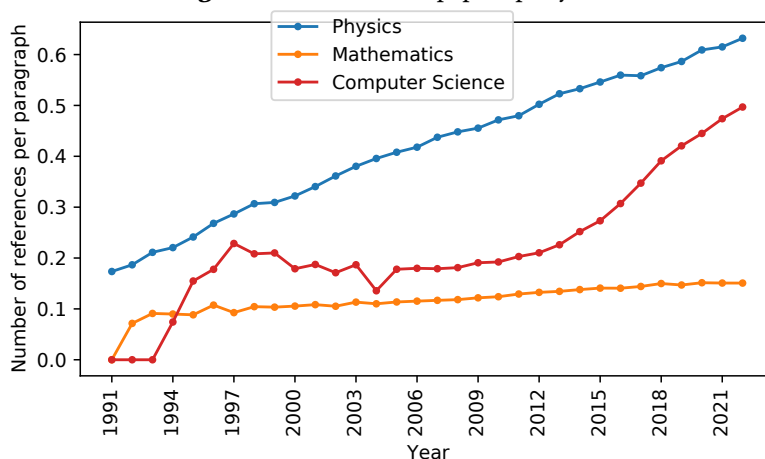


Figure 4.5.: Reference density per year.

and the development of models for claim verification [191], document retrieval [151], summarization [126], or information extraction [190].

Due to its similarities in structure and contained information, unarXive 2022 is equally suitable for the applications named above. Beyond these, we provide data for two NLP tasks on unarXive 2022, namely content based citation recommendation and IMRaD classification, which are described in the following.

Content Based Citation Recommendation Given a piece of text and a citation-marker position, the task of content based citation recommendation entails identifying publications which are suitable to cite in the given text at the given position [28, 58]. Large full-text corpora of publications with a citation network provide a rich source for supervision of machine learning (ML) models for this task. That is, human made citations are used as training examples, or for evaluating models in a citation re-prediction setting. From the permissively licensed papers in our data set we use all in-text citation markers with a linked reference cited at least three times, to allow splitting into train, dev, and test

data. The result is 2.5 M items consisting of (1) a paragraph and citation marker position (model input), and (2) the ID of the cited document (desired model output).

IMRaD Classification Scientific publications are usually structured into sections commonly summarized as “Introduction, Methods, Results, and Discussion” (IMRaD). Classifying sections of scientific text into these four classes is done, for example, in fine-grained citation classification. Because conventions differ between disciplines, we prepare data for this task for computer science papers only. To aforementioned four classes we add the common “Related Work” section as a fifth class. From the permissively licensed computer science papers in our data, we use those that are unambiguously assignable to one of the five classes. The result is 530 k items consisting of (1) the paragraph text (model input), and (2) the class (desired model output). An exemplary application scenario for a model trained on this data is a paper writing assistant that can detect parts in a manuscript, which might be better placed in a different section (e.g. discussion rather than results).

4.3.4.3. Distribution

Under consideration of the FAIR principles, we chose the following well established distribution channels and licenses for our data set, aforementioned NLP task data, as well as our source code.

- The **data set** is distributed on Zenodo.
 - <https://doi.org/10.5281/zenodo.7752615> [last accessed: 2023-11-10] (open subset)
 - <https://doi.org/10.5281/zenodo.7752754> [last accessed: 2023-11-10] (full)
 In accordance with the licensing terms of our source data, we share our data set in two versions.
 - (1) The subset generated from permissively licensed source data (165 k publications, 9%) is openly accessible.
 - (2) The full data set, generated partially from source data under arXiv.org’s “non-exclusive license to distribute,”¹⁸ is accessible through Zenodo’s “restricted access” policy,¹⁹ making it possible to grant access to the data on request given the intended use is in accordance with the license terms.
- The **NLP task data** is provided on the Hugging Face Hub.
 - https://hf.co/datasets/saier/unarXive_citrec [last accessed: 2023-11-10]
 - https://hf.co/datasets/saier/unarXive_imrad_clf [last accessed: 2023-11-10]
 This facilitates easy access and use by the NLP community.
- The **source code** for creating the data set is shared on GitHub under the MIT License.
 - <https://github.com/Il1Depence/unarXive> [last accessed: 2023-11-10]
 Sharing the code openly and permissively licensed allows anyone to freely modify

¹⁸ See <http://arxiv.org/licenses/nonexclusive-distrib/1.0/> [last accessed: 2023-11-10].

¹⁹ See <https://about.zenodo.org/policies/> [last accessed: 2023-11-10].

and extend the code to their needs. This makes, for example, integration into other NLP projects such as benchmarks and frameworks possible.

4.3.5. Conclusion

We propose unarXive 2022, a data set generated from 1.9 M \LaTeX paper sources and suitable for a wide variety of analyses and NLP applications. We base our approach to data set creation and format on existing works, while also addressing their shortcomings. Improving upon these tried and tested predecessors, unarXive 2022 offers the most complete citation network and most structured content compared to existing data sets, and is surpassed in size only by the PMC OAS, which covers a different set of disciplines.

Together with our data set, we provide data for two NLP tasks, content based citation recommendation and IMRaD classification, to facilitate its usage. We furthermore distribute our work under consideration of the FAIR principles, sharing it through well established channels and permissively licensed, thereby ensuring proper accessibility, easy use, and possibilities for adaption and extension.

We plan to incrementally update our data set with new arXiv.org submissions. For future developments, we note the importance of mathematical notation in academic publications, as reflected by recent SemEval tasks in 2021 and 2022 [78, 107]. Similar to existing projects,²⁰ we plan to investigate novel analyses and applications based on the combination of our data set’s citation network and structured mathematical notation.

Author Contributions

Tarek Saier: Conceptualization, Data curation (lead), Formal analysis, Methodology, Software (lead), Visualization, Writing – original draft, Writing – review & editing. Johan Krause: Data curation (support), Software (support). Michael Färber: Writing – review & editing.

Acknowledgements

This work was partially supported by the German Federal Ministry of Education and Research (BMBF) via [KOM,BI], a Software Campus project (01IS17042). The authors acknowledge support by the state of Baden-Württemberg through bwHPC. We thank Johannes Reber for supporting early stages of the software development.

²⁰ See <https://github.com/PierreSenellart/theoremkb> [last accessed: 2023-11-10].

4.4. Result Assessment

The work in this chapter addresses the following research task.

- ❖ **RT2:** *Citation Network Completeness* - develop a method to link literature references, that is able to link more references than are linked in existing data sets, while not compromising on link correctness or processing efficiency.

With the inter-reference blocking and matching method presented in Section 4.2, we achieve a 90% increase in linked references, a five-fold increase in bibliographic couplings, and a nine-fold increase in connected in-text citation markers. Furthermore, with the improved corpus creation method presented in Section 4.3, we achieve a new state-of-the-art reference matching success rate of 44.4%. Different from our contributions to ❖ **RT2** in Chapter 3, we are here able to make direct comparisons to related work. Our reference matching success rate of 44.4% compares favorably to our previous method (42.6%), S2ORC (31.1%), and arXMLiv (no citation network). Accordingly, we deem ❖ **RT2** successfully achieved.

In terms of the overarching research goal of enabling higher-quality scholarly data (see Table 2.4 in Chapter 2), the work presented in this chapter makes the following contributions.

Scholarly Data Quality Contributions - [2, 3]	
Crit.	Contribution
Rel_{SDR}	Mathematical notation included in the structured document representation
$\text{Tim}_{\text{C/S}}$	Publications included up until end of the most recent full year
Coy_{CN}	OpenAlex, arXiv, and PubMed IDs as well as DOIs included
Coy_{SDR}	Section structure and section type information included
Cos_{CN}	Blocking and matching procedure enabling increase in linked references and bibliographic couplings
Cos_{SDR}	Full-text of documents as well as figure and table captions included

Rel_{SDR} Our improved document conversion methodology provides structured mathematical notation as part of our document representations. The resulting data is therefore of high relevance for use cases focussing on phenomena manifested within them, such as mathematical information retrieval [107] and the analysis of theorems.²¹ Our data source being arXiv, which has a large coverage of physics, mathematics, and computer science papers, additionally is beneficial in this regard, as all three are disciplines, where mathematical notation is comparably relevant and frequently used.

²¹ See <https://github.com/PierreSenellart/theoremkb> [last accessed: 2024-02-04].

Tim_{C/S} We apply our method for generating scholarly data on all documents on arXiv until end of the most recent full year. Accordingly, the resulting corpus contains more recent documents than data sets released earlier.

Co_{YCN} By linking references to OpenAlex, we are able to provide a large number of document identifiers in our citation network. In particular, these are OpenAlex IDs, arXiv IDs, PubMed IDs, and DOIs. This makes our data a versatile target for comparing and combining it with other data.

Co_{YSDR} Our document representation provides a section and paragraph structure as well as paragraph type information (text, listing, etc.). This enables fine-grained document content comparison, filtering, and combination with external data.

Co_{SCN} With our inter-reference blocking and matching method, we provide a means to improve citation network completeness through increased linked references (Section 4.2). Furthermore, our improved corpus creation method achieves a new state-of-the-art reference matching success rate of 44.4% (Section 4.3).

Co_{SDR} In addition to papers' full-text content, we provide figure and table captions linked within the text. This makes our structured document representations more complete than before.

5

References Across Languages

This chapter is based on the following publications.



Tarek Saier and Michael Färber. “A Large-Scale Analysis of Cross-lingual Citations in English Papers”. In: *Digital Libraries at Times of Massive Societal Transition*. Springer International Publishing, 2020, pp. 122–138. ISBN: 978-3-030-64452-9. DOI: 10.1007/978-3-030-64452-9_11



Tarek Saier, Michael Färber, and Tornike Tsereteli. “Cross-Lingual Citations in English Papers: A Large-Scale Analysis of Prevalence, Usage, and Impact”. In: *International Journal on Digital Libraries* 23.2 (June 2022), pp. 179–195. ISSN: 1432-1300. DOI: 10.1007/s00799-021-00312-z

The work in this chapter addresses the following research task.

- ❖ **RT3: Inclusion of Non-English Publications** - find and implement an approach to include non-English publications into a large-scale, high-quality scholarly data set.

5.1. Overview

In this chapter, we address the scholarly data limitation of disregarding non-English publications. Our corpus, *unarXive*, while not restricted to any language, predominantly consists of English publications. This is simply because the publications that authors submit to arXiv.org happen to be mostly English. However, what the corpus is suitable for studying is *citations* to non-English publications.

For this, we find a method to reliably identify cross-lingual citations in English publications. Based on this, we conduct the so far largest analysis of this type of citation, covering over one million publications. We analyze cross-lingual citations’ prevalence, usage, as well as impact, and identify trends over time as well as challenges.

At the end of the chapter, in Section 5.7, we assess the achievement of the research task, as well as the contributions made in terms of the overarching research goal of enabling higher-quality scholarly data.

5.2. Introduction

Citations are an essential tool for scientific practice. By allowing authors to refer to existing publications, citations make it possible to position one’s work within the context of others’, critique, compare, and point readers to supplementary reading material. In other words, citations enable scientific discourse. Because of this, citations are a valuable indicator for the academic community’s reception of and interaction with published works. Their analysis is used, for example, to quantify research output [82], qualify references [14], and detect trends [37]. Furthermore, citations can be utilized to aid researchers through, for example, summarization [55] or recommendation [125, 58] of papers, and through applications driven by document embeddings in general [46].

Because these analyses and applications require data to be based on, the availability of citation data or lack thereof is decisive with regard to the areas in which respective insights can be gained and approaches developed. Here, the literature points in two major directions of lacking coverage—namely the humanities [48, 96] and non-English publications [189, 121, 137, 142, 127]. Because most large scholarly data sets are either artificially limited to few languages (e.g., English only) or do not provide language metadata, a particular practice not well researched so far is cross-lingual citation. That is, references where the citing and cited documents are written in different languages (see (vi) in Figure 5.1). Cross-lingual citations are, however, important bridges between otherwise insufficiently connected “language silos” [170, 142].

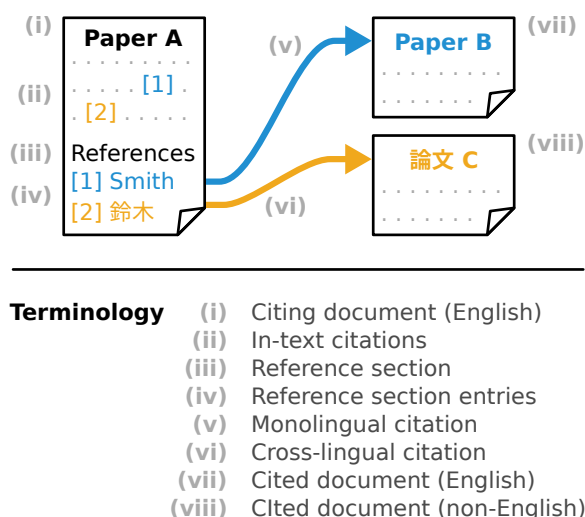


Figure 5.1.: Schematic explanation of terminology.

English currently being the de facto academic lingua franca [140], citations from non-English languages to English are significantly more prevalent than the other way around. This dichotomy is reflected in existing literature, where usually either citations from English [96, 117], or to English [181, 89, 90, 167] are analyzed. As both directions involve a non-English document on one side of the citation, the analysis of either is challenging with today’s anglocentric state of citation data.

Setting our focus to cross-lingual citations *from English*, we perform a large-scale analysis on over one million documents. In line with existing literature, we determine the prevalence of cross-lingual citations across multiple dimensions. Additionally, we investigate the citations’ usage as well as impact. In particular, the following questions are addressed.

- How prevalent are English to non-English references? We consider prevalence in general, in different disciplines, across time, and within publications that use them (Section 5.5.1, “Prevalence”).
- In what circumstances are cross-lingual citations in English papers used? Here we consider self-citation, geographic origin, as well as citation function and sentiment (Section 5.5.2, “Usage”).
- What is the impact of cross-lingual citations in English documents? We consider the aspects of acceptance, data mining challenges, as well as impact on the success of a publication (Section 5.5.3, “Impact”).

Through our analysis, we make the following contributions.

1. We conduct an analysis of cross-lingual citations in English papers that is considerably more extensive than existing literature in terms of corpus size as well as covered languages, time, and disciplines. This not only makes the results more representative of the areas covered, but also enables the use of our collected data for machine learning based applications such as cross-lingual citation recommendation.
2. We propose an easy and reliable method for identifying cross-lingual citations from English papers to publications in non-Latin script languages (e.g., Russian and Chinese).
3. We highlight key challenges for handling cross-lingual citations that can inform future developments in scholarly data mining.
4. To facilitate further research, we make our collected data, source code, and full results publicly available.¹

The remainder of this chapter is structured as follows. After briefly addressing our use of terminology down below, we give an overview of related work in Section 5.3. In Section 5.4 we discuss the identification of cross-lingual citations, data sources considered, and our data collection process. Subsequent analyses of the identified citations are presented in

¹ See <https://github.com/Il1Depence/cross-lingual-citations-from-en> [last accessed: 2023-11-10].

Table 5.1.: Comparison of corpora.

Work	Type ^a	#Docs	#References	#Years	#Disciplines
Kellsey and Knieval [96]	en→*	468	16k	5 ^b	4
Lillis et al. [117]	en→*	240	10k	7	1
Schrader [167]	*→en	403	5k	2	1
Tang et al. [181]	zh→en	2k	17k	10	1
Jiang et al. [89, 90]	zh→{en,zh}	14k	38k	n/a	1
Kirchik et al. [99]	{en,ru}→ru	497k	n/a	17	(unrestricted)
Ours	en→*	1.1M	39M	27	3

^a type=focus reference type (en=English, ru=Russian, zh=Chinese, *=any)

^b over a span of 40 years

Section 5.5. We end with a discussion of our findings and concluding remarks in Section 5.6, and an overarching result assessment in Section 5.7.

Terminology

Because *citation*, *reference* and related terms are not used consistently in literature, we briefly address their use in this chapter. As shown in Figure 5.1, a *citing* document creates a bibliographical link to a *cited* document. We use the terms *citation* and *reference* interchangeably for this type of link (e.g. “(vi) in Figure 5.1 marks a cross-lingual reference” or “Paper A makes two citations”). The textual manifestation of a bibliographic reference, often found at the end of a paper (e.g. “[1] Smith” in Figure 5.1), is referred to as *reference section entry*, or sometimes *reference* for short. We call the combined set of these entries *reference section*. Lastly, parts within the text of a paper, which contain a marker connected to one of the reference section entries, are called *in-text citations*.

5.3. Related Work

Existing literature on cross-lingual citations in academic publications covers analyses as well as approaches to prediction tasks. These are, however, only based on small corpora or restricted to specific language pairs. As shown in Table 5.1, our work is based on a considerably larger corpus which is also more comprehensive in terms of the time span and disciplines that are covered.

In the following, we describe the works in Table 5.1 in more detail, reporting on the key corpus characteristics and findings. This is complemented by a short overview of existing literature on various types of cross-lingual interconnections in media other than academic publications.

5.3.1. Cross-Lingual Citations in Academic Publications

Literature concerning cross-lingual citations in academic publications can be found in the form of analyses and applications. In [96] Kellsey and Knievel conduct an analysis of 468 articles containing 16,138 citations. The analysis spans 4 English language journals in the humanities (disciplines: history, classics, linguistics, and philosophy) over 5 particular years (1962, 1972, 1982, 1992, and 2002). They count cross-lingual citations to English, German, French, Italian, Spanish, Portuguese, and Latin, while further languages are grouped into a category “other.” The authors find that 21.3% of the citations in their corpus are cross-lingual, but note strong differences between the covered disciplines. Over time, they observe a steady total, but declining relative number of cross-lingual citations per article. The authors furthermore find, that the ratio of publications that contain at least one cross-lingual citation is increasing.

Lillis et al. [117] investigate if the global status of English is impacting the “citability” of non-English works in English publications. They base their analysis on 240 articles from 2000 to 2007 in psychology journals, and furthermore use the Social Sciences Citation Index and ethnographic records. Their corpus contains 10,688 references, of which 8.5% are cross-lingual. Analyzing the prevalence of references in various contexts, they find that authors are more likely to cite a “local language” in English-medium national journals than in international journals. Further conducting analyses of, for example, in-text citation surface forms, they come to the conclusion that there are strong indicators for a pressure to cite English rather than non-English publications.

Similar observations are made by Kirchik et al. [99] concerning citations to Russian. Analyzing 498,221 papers in Thomson Reuters’ Web of Science between 1993 and 2010, they find that Russian scholars are more than twice as likely to cite Russian publications when publishing in Russian language journals (21% of citations) than when they publish in English (10% of citations).

In [167] Schrader analyzes citations from non-English documents to English articles in open access and “traditional” journals. The corpus used comprises 403 cited articles published between 2011 and 2012 in the discipline of library and information science. The articles were cited 5,183 times (13.8% by non-English documents). In their analysis the author observes that being open access makes no statistically significant difference for the ratio of incoming cross-lingual citations of an article, or the language composition of citations a journal receives.

Apart from analyses, there are also approaches to prediction tasks based on cross-lingual citations [181, 89, 90, 125]. Tang et al. [181] propose a bilingual context-citation embedding algorithm for the task of predicting suitable citations to English publications in Chinese sentences. To train and evaluate their approach, they use 2,061 articles from 2002 to 2012 in the Chinese Journal of Computers, which contain citations to 17,693 English publications. Comparing to several baseline methods, they observe the best performance for their novel system. Similarly, in [89] and [90] Jiang et al. propose two novel document embedding methods jointly learned on publication content and citation relations. The corpus used in

both cases consists of 14,631 Chinese computer science papers from the Wanfang digital library. The papers contain 11,252 references to Chinese publications and 27,101 references to English publications. For the task of predicting a list of suitable English language references for a Chinese query document, both approaches are reported to outperform a range of baseline methods.

5.3.2. Cross-Lingual Interconnections in Other Types of Media

Apart from academic publications, cross-lingual connections are also described in other types of media. Hale [77] analyzes cross-lingual hyperlinks between online blogs centered around a news event in 2010. In a corpus of 113,117 blog pages in English, Spanish, and Japanese, 12,527 hyperlinks (5.6% of them cross-lingual) are identified. Analysis finds that less than 2% of links in English blogs are cross-lingual, while the number in Spanish and Japanese blogs is slightly above 10%. Hyperlinks between Spanish and Japanese are almost non-existent (7 in total). Further investigating the development of links over time, the author observes a gradual decrease in language group insularity driven by individual translations of blog content—a phenomenon described as “bridgeblogging” by Zuckerman [207]. Similar structural features are reported by Eleta et al. [54] and Hale [76] for Twitter, where multilingual users are bridging language communities.

Focusing on types of information diffusion that are not textually manifested through connections such as bibliographic references and hyperlinks, there also is literature on cross-lingual phenomena on collaborative online platforms, such as the study of cross-lingual information diffusion on Wikipedia [98, 165].

Lastly, as with academic publications, there furthermore exists literature on link prediction tasks. In [91] Jin et al. analyze cross-lingual information cascades and develop a machine learning approach based on language and content features to predict the size and language distribution of such cascades.

5.4. Data Collection

In this section, we first discuss how to identify cross-lingual citations. Subsequently, we outline the steps of data source selection and corpus construction. Lastly, we describe the key characteristics of our corpus.

5.4.1. Identification of Cross-Lingual Citations

Identifying cross-lingual citations requires information about the language of the citing and cited document. However, this is often missing in scholarly data sets.² Identifying the

² Details are provided in Section 5.4.2.

involved documents’ language when it is not given in metadata, however, is challenging, because (a) the full-text, especially of the cited documents, is not always available, (b) abstracts are not reliable because non-English publications often provide an additional English abstract, and (c) language identification on short strings (e.g., titles in references) does not achieve sufficient results with existing techniques [88].

To nevertheless be able to conduct an analysis of cross-lingual citations on a large scale, we utilize the common practice of authors appending an explicit marker in the form of “(in <Language>)” to such references. This shifts the requirements from language metadata or language identification to the existence of reference section entries in the data. This is because the language of the cited document is given by the “<Language>” part of the marker, and the language the marker itself is written in (i.e., English) provides the citing document’s language. For example, the reference section entry “*M. Saitou, ‘Hydrodynamics on non-commutative space’ (in Japanese), [...]*”³ by itself contains enough information to determine that the cited document is written in Japanese and the citing document is written in English.

The question then remains, how common the practice of using such explicit markers is—that is, to cite, for example, “*A Modern Model Description of Magnetism (in Russian)*” instead of “*Современное модельное описание магнетизма*”.⁴ To answer this question, we perform a preliminary analysis on the data set unarXive [1], which comprises 39 million reference section entries. Specifically, we conduct a large automated analysis on all reference section entries in the data set and additionally perform a smaller, manual analysis on a stratified sample of 5,000 references.

In the large automated analysis, we first identify the cited document’s title within references using the state-of-the-art [185] reference string parser module of GROBID [123], and then determine the title’s language using the language identification tool Lingua,⁵ which is specialized for very short text. Manually inspecting our results, we note that non-Latin script languages (e.g., Chinese, Japanese, Russian) are detected reliably,⁶ but Latin script languages (e.g., German and French) are not. For instance, many English titles are falsely identified as German.

For non-Latin script languages, which we show in Table 5.2, only a small fraction of cross-lingual citations is not explicitly marked. We observe ratios of unmarked cross-lingual citations relative to explicit markers consistently below 2%.⁷

³ Found in arXiv:1612.01831.

⁴ Referring to arXiv:1103.5123.

⁵ See <https://github.com/pemistahl/lingua> [last accessed: 2023-11-10].

⁶ To be more precise, no language that uses a script different to the Latin alphabet appears to be falsely identified as English. We are, however, not able to judge whether languages using the same non-Latin script—such as languages written in Cyrillic—are distinguished correctly by Lingua.

⁷ Because our analysis is based on language identification of the titles of cited publications, we cannot detect when a non-English work is cited with a translated title *and* no explicit language marker.

Table 5.2.: References to non-Latin script languages in the automated analysis.

Cited Language	#marked	#unmarked
Russian	23,922	303 (1.3%)
Chinese	2,351	10 (0.4%)
Japanese	1,843	5 (0.3%)
Ukrainian	876	15 (1.7%)
Bulgarian	67	0 (0.0%)
Greek	60	1 (1.7%)

Table 5.3.: Results of manual labeling.

Cited Language	#references	#marked
(n/a) ^a	2,737	0
English	2,188	0
French	33	1
German	27	0
Russian	8	6 ^b
Italian	5	1
Chinese	1	1
Japanese	1	1

^a These references did not contain the title of the cited document, which is common in physics papers.

^b The two remaining unmarked references contained the cited publication's title only transliterated into the Latin alphabet.

To get a reliable estimate for Latin script languages as well, we additionally perform a smaller, manual analysis. To this end, we label a stratified sample⁸ of 5,000 references from unarXive with the reference's language as well as whether an explicit language marker was used or not. The results of our evaluation are shown in Table 5.3. In accordance with our automated large analysis, we observe that non-Latin script languages are generally explicitly marked. For Latin script languages, however, explicit marking appears to be considerably less common. We additionally evaluate the automated language identification results for our manually annotated references and measure F1 scores of 0.48, 0.46, and 0.60 for French, German, and Italian respectively. Notably, less than half of the references with German titles are detected (44% recall) and more than half of the references identified as German are false positives (48% precision).

The results of above preliminary investigations have two consequences for the findings in our main analyses, which are based on explicit language markers. First, a direct

⁸ The sample was stratified according to the referencing document's discipline and month of publication.

Table 5.4.: Overview of data sets.

Data set	#Docs	Lang. Meta ^a	Refs. to ^b	Reference sections	Used
MAG ^c [173, 192]	230M	(48% ^d)	MAG	-	✓
CORE ^e	123M	1.79%	CORE	-	
S2ORC [122]	81M	-	S2ORC	34% (in GROBID parse)	
PMC OAS ^f	2M	-	mixed	100% (in JATS XML)	
unarXive [1]	1M	-	MAG	100% (dedicated entity)	✓

^a Language metadata

^b References resolved to

^c Using version 2019-12-26

^d Language given for source URLs (not always matching paper language)

^e See <https://core.ac.uk/> [last accessed: 2023-11-10]. Using version 2018-03-01

^f See <https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/> [last accessed: 2023-11-10]

comparison between our results on non-Latin and Latin script languages is only valid for *explicitly marked* cross-lingual citations, as there is a notable amount of undetected cross-lingual citations for Latin script languages. Second, the number of undetected cross-lingual citations for non-Latin script languages such as Chinese, Japanese, and Russian, is negligible. Accordingly, concerning these languages, our results are valid for cross-lingual citations *regardless of language markers*.

5.4.2. Data Source Selection

As our data source we considered five large scholarly data sets commonly used for citation related tasks [97, 58]. Table 5.4 gives an overview of their key properties. The Microsoft Academic Graph (MAG) and CORE are both very large data sets with some form of language metadata present. In the MAG the language is given not for documents themselves, but for URLs associated with papers. CORE contains a language label for 1.79% of its documents. S2ORC, the PubMed Central Open Access Subset (PMC OAS), and unarXive do not offer language metadata, but all contain some form of reference sections (GROBID output, JATS [84] XML, and raw strings extracted from \LaTeX source files respectively).

From these five, we decided to use unarXive and the MAG. This decision was motivated by two key reasons: (1) metadata of cited documents, and (2) evaluation of the acceptance of cross-lingual citations in English papers. As for (1), both S2ORC and the PMC OAS link references in their papers to document IDs within the data set itself (only partly in the PMC OAS, where also MEDLINE IDs and DOIs are found [70]). This is problematic in our case, because S2ORC is restricted to English papers, and the PMC OAS is constrained to Latin script contents,⁹ which means metadata on non-English cited documents is non-existent (S2ORC) or very limited (PMC OAS). In unarXive, on the other hand, references are linked

⁹ See <https://www.ncbi.nlm.nih.gov/pmc/about/faq/#q16> [last accessed: 2023-11-10].

to the MAG, which contains metadata on publications regardless of language. Concerning reason (2), the fact that unarXive is built from papers on the preprint server arxiv.org, and the MAG contains metadata on paper’s preprint *and* published versions, allows us to analyze whether or not cross-lingual citations are affected by the peer review process.

With these two data sources selected, the extent of our analysis is over one million documents, across 3 disciplines (physics, mathematics, computer science), over a span of 27 years (1992–2019).

5.4.3. Data Collection

To identify references with “(in <Language>)” markers, we iterate through the total of 39.7M reference section entries in unarXive and first filter for the regular expression $\backslash(\backslash s^*in\backslash s^*[a-zA-Z][a-z]+\backslash s^*\backslash)$. This yields 51,380 matches with 207 unique tokens following “in” within the parentheses. Within these 207 tokens we manually remove those referring to non-languages (e.g., “press” or “preparation”) and correct misspellings (e.g., “japanease” or “russain”), resulting in 44 unique language tokens. These are (presented in ISO 639-1 codes) be, bg, ca, cs, da, de, el, en, eo, es, et, fa, fi, fr, he, hi, hr, hu, hy, id, is, it, ja, ka, ko, la, lv, mk, mr, nl, no, pl, pt, ro, ru, sa, sk, sl, sr, sv, tr, uk, vi, and zh. These 44 languages cover 43 of the 78 languages, in which journals indexed in the Directory of Open Access Journals¹⁰ (DOAJ) are published as of July 2020. The one language found in our data, but with no journal in the DOAJ, is Marathi. In terms of journal count by language, above 44 languages cover 97.54% of the DOAJ. In total, our data contains 33,290 reference section entries in 18,171 unique citing documents. We refer to this set of documents as the *cross-lingual set*.

To analyze differences between papers containing cross-lingual citations in unarXive and a comparable random set, we also generate a second set of papers. To ensure comparability we go through each year of the cross-lingual set, note the number of documents per discipline and then randomly sample the same number of documents from all of unarXive within this year and discipline. This means the *cross-lingual set* and the *random set* have the same document distribution across years and disciplines. Table 5.5 gives an overview of the resulting data used.

5.5. Results

In the following, we describe the results of our analyses with regard to the questions laid out in the introduction. We begin with general numbers concerning the *prevalence* of cross-lingual citations. These results are based on unarXive alone. This is followed by more in-depth observations regarding cross-lingual citations’ *usage* (e.g., the underlying motivation

¹⁰ See <https://doaj.org/> [last accessed: 2023-11-10].

Table 5.5.: Overview of data used.

	Cross-lingual set	Random set	unarXive
#Docs	18,171	18,171	1,192,097
#Docs (MAG)	16,300	16,464	1,087,765
#Refs	635,154	536,672	39,694,083
#Refs (MAG)	290,421	242,090	15,954,664
#Cross-lingual refs	33,290	642	33,290

*docs = documents,
 refs = reference section entries,
 (MAG) = with a MAG ID.

Table 5.6.: Most prevalent languages.

Language	#References	#Documents
Russian	23,922	12,304
Chinese	2,351	1,582
Japanese	1,843	1,397
German	1,244	965
French	931	719

or the citation’s function) and *impact* (e.g., acceptance by reviewers or challenges for data mining). These subsequent in-depth analyses additionally utilize the MAG metadata.

5.5.1. Prevalence

We find “(in <Language>)” markers in 33,290 out of 39,694,083 reference section entries (0.08%). These appear in 18,171 out of 1,192,097 documents (1.5%)—in other words in every 66th document. Of these 18k documents, 17,223 cite one language other than English, 864 cite two, 76 three, 7 documents four, and a single document cites works in English and five further languages (Russian, French, Polish, Italian, and German). The five most common language pairs within a single document are Russian-Ukrainian (277 documents), German-Russian (166), French-Russian (135), French-German (68), and Chinese-Russian (59).

Table 5.6 shows the absolute number of reference section entries and unique citing documents for the five most prevalent languages, which combined make up over 90% in terms of both references and documents. As we can see, Russian is by far the most common, making up about two thirds of the cross-lingual set. When breaking down these numbers by year or discipline, it is important to also factor in the distribution of documents along these dimensions in the whole data set. Doing so, we show in Figure 5.2 the relative number of documents with cross-lingual citations over time for each of the aforementioned five languages. While the numbers in earlier years can be a bit unstable due to low numbers of total documents, we can observe a downwards trend of citations to Russian,

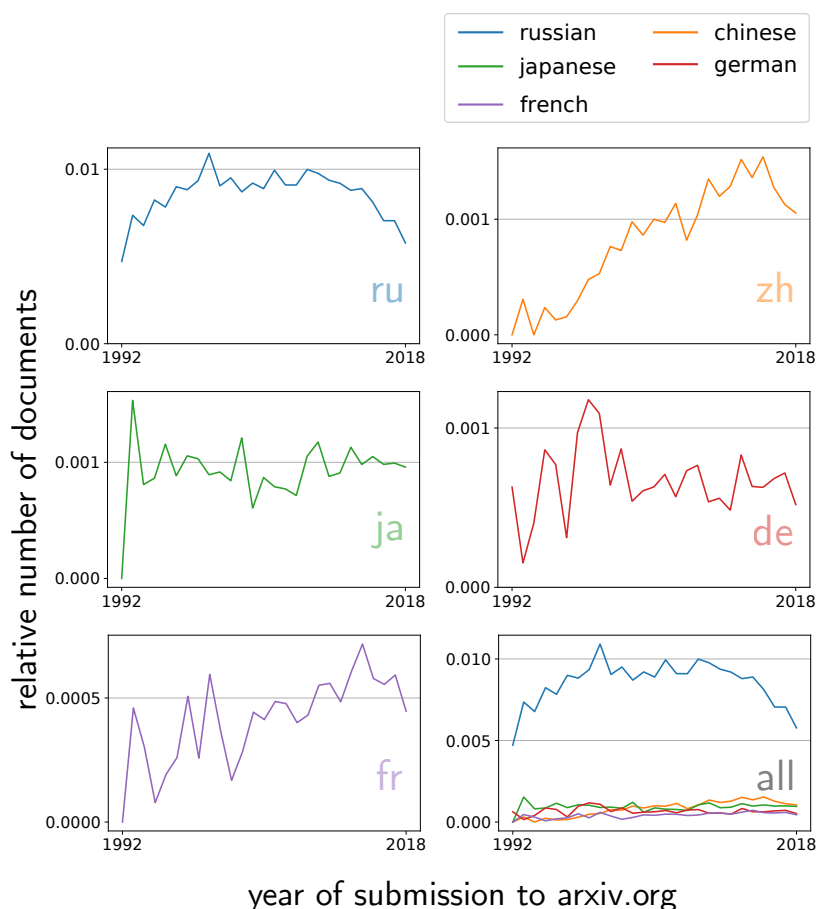


Figure 5.2.: Relative number of documents citing Russian, Chinese, Japanese, German, and French works. Showing all aforementioned in the bottom right.

an upwards trend of citations to Chinese, and a somewhat stable proportion in documents citing Japanese works. Looking at the numbers per discipline in Figure 5.3, we can see that cross-lingual citations occur most often in mathematics papers, and are about half as common in physics and computer science.

Lastly, within the reference section of a document that has at least one cross-lingual citation, the mean value of “cross-linguality” (i.e., what portion of the reference section is cross-lingual) is 0.083 with a standard deviation of 0.099. Breaking these numbers down by discipline, we can see in Figure 5.4 that there is no large difference, although mathematics papers tend to have a slightly higher portion of cross-lingual citations. The mean values for mathematics, physics and computer science are 0.090, 0.078, and 0.080 respectively.

Regarding prevalence, we observe that in English papers in the disciplines of physics, mathematics, and computer science about 1 in 66 publications contains at least one explicitly marked citation to a non-English document. About two thirds of these citations are to Russian documents, although in the last years there is a downwards trend with regard to Russian and an upwards trend in citations to Chinese. Furthermore, cross-lingual citations appear about twice as often in mathematics compared to physics and computer science.

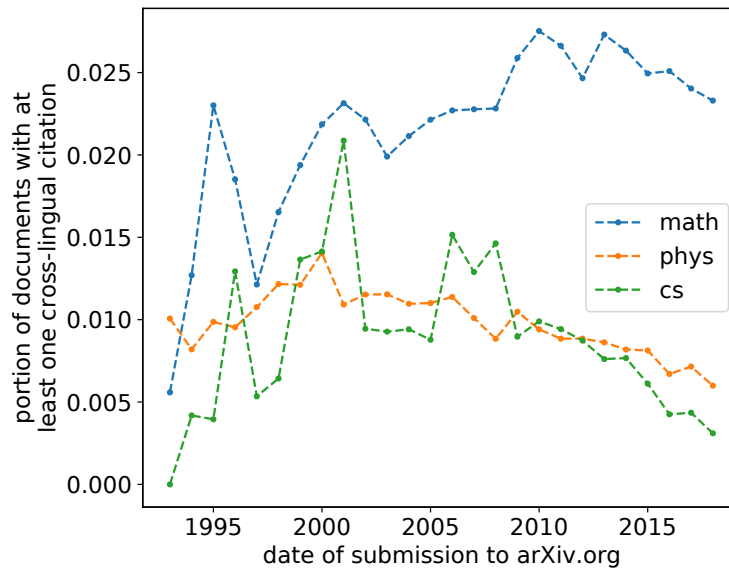


Figure 5.3.: Relative number of mathematics, physics, and computer science documents citing non-English works.

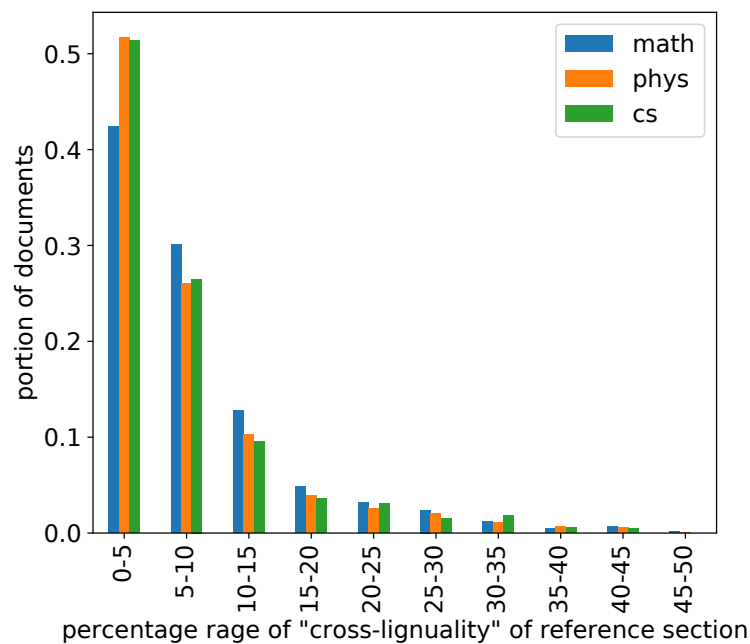


Figure 5.4.: "Cross-linguality" of reference sections by discipline.

These observations suggest that while cross-lingual citations are not very frequent in general, they might be worth considering in applications dealing with specific disciplines and languages (e.g. citations to Russian in mathematics publications).

Table 5.7.: Self-citations.

References to	Self-citations	
	loose	strict
non-English	19%	5%
English	17.9%	11.3%

5.5.2. Usage

Regarding the usage of cross-lingual citations in English publications we analyze four different aspects. (1) Whether or not self-citations are a driving factor, (2) to what degree the geographical origin of a cross-lingual citation is correlated with the cited document’s language, (3) what function they serve, and (4) what sentiment they express toward the cited document.

5.5.2.1. Self-citation

To assess the relative degree of self-citation when referring to publications in other languages, we compare the ratio of self-citations in (a) the *cross-lingual citations* within the documents of the cross-lingual set, and (b) the *monolingual citations* within the documents of the cross-lingual set. Comparing two sets of citations from identical documents allows us to control for confounding effects such as author specific self-citation bias.

To determine self-citation, we rely on the author metadata in the MAG and therefore require both the citing and cited document of a reference to have a MAG ID. Within the cross-lingual set, this is the case for 3,370 cross-lingual references and 264,341 monolingual references. While at first, we strictly determine a self-citation by author IDs in the MAG being identical, manual inspection of matches and non-matches reveals, that author disambiguation within the MAG is somewhat lacking—that is, in a non-trivial amount of cases there are several IDs for a single author. We therefore measure self-citation by two metrics. A strict metric which only counts a match of MAG IDs, and a loose metric which counts an overlap of the sets of author names on both ends of the reference as a self-citation.

Table 5.7 shows that going by the strict metric, self-citation is twice as common in monolingual citations. Applying the loose metric, however, self-citation appears to be slightly more common in cross-lingual citations. The larger discrepancy between the results of the strict and loose metric for cross-lingual citations suggests that authors publishing in multiple languages might be less well disambiguated in the MAG. With regard to self-citation being a motivating factor for cross-lingual citations—be it, for example, due to the need to reference one’s own prior work—, we can note that our data does not suggest this to be the case. Authors using cross-lingual citations appear to be at least equally as likely to self-cite when referencing English works.

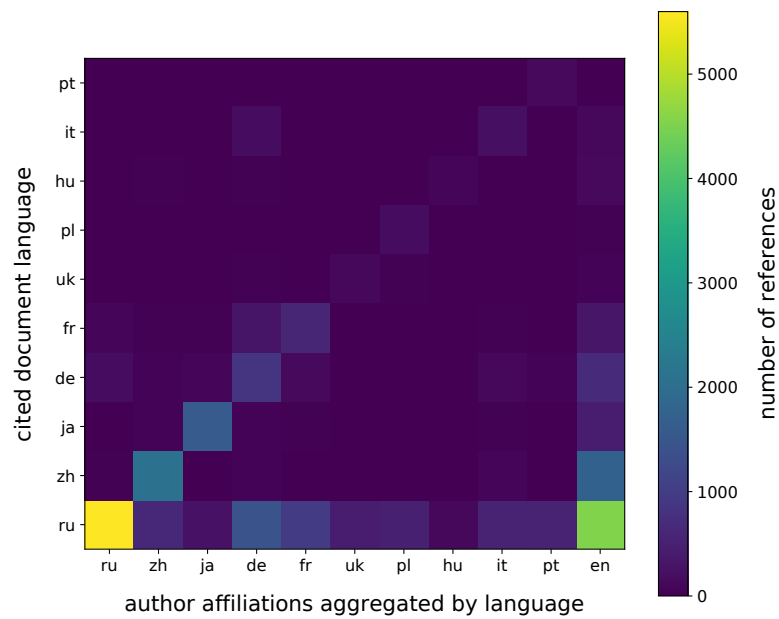


Figure 5.5.: Geographic origin of cross-lingual citations to the ten most cited languages (absolute count).

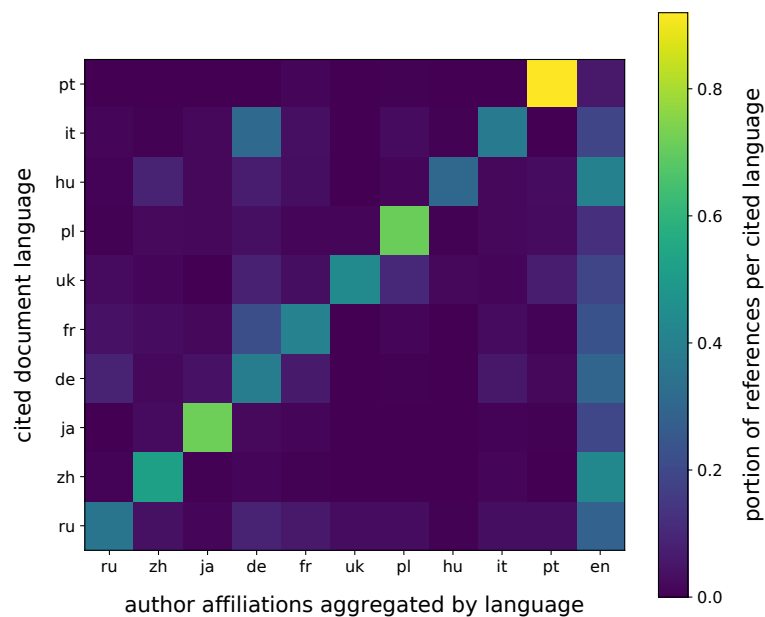


Figure 5.6.: Geographic origin of cross-lingual citations to the ten most cited languages (relative count).

5.5.2.2. Geographical Origin

In this section we analyze the geographical origin of cross-lingual citations. As a measure for geographical origin we use the country in which a citing author’s affiliation is located. We refer to a citation as being to a “local language” or of “local origin”, if the cited document’s language is the most commonly spoken language in the affiliation’s location.

An example of this would be a researcher affiliated with a research institution located in Russia, being the author of a paper in which they cite a publication written in Russian.

For our analysis, we rely on author affiliation metadata in the MAG. We start off with all documents in the cross-lingual set that have a MAG ID.¹¹ From those, we select all which provide information on the authors' affiliations.¹² This leaves us with 7,522 out of 16,300 papers. To associate an author's affiliation with a language, we use the most commonly spoken language in the country or territory.¹³ Grouping affiliations by language, we can then view the correlation of (a) cited languages and (b) language grouped affiliations in two ways. On the one hand, we can see for each cited language how many of the citations are of local origin—compared to, for example, from an English-speaking country. On the other hand, we can see for each language group of affiliations how many cross-lingual citations are to a local language. Our results of this analysis are shown for the 10 most commonly cited languages in Figures 5.5 and 5.6, and for all identified cited languages in Appendix A.1.

Figure 5.5 shows citation numbers in absolute terms. Looking, for example, at citations to Russian publications (the bottom row of the figure), we can see that the largest amount of citations originates from Russian-speaking countries (5,599 out of 18,672) followed by English-speaking countries (4,535) and German-speaking countries (1,427).

In Figure 5.6 we show relative numbers per cited language. That is, the values of each row add up to 1. Here we can see that citations to Japanese, Polish and particularly Portuguese appear to be of local origin comparatively often, with 68% for Japanese, 64% for Polish and 86% for Portuguese. Overall we observe that cross-lingual citations are most often either of local origin or from an English-speaking country. Evaluated over all languages, 37% of cross-lingual citations are local (the diagonal in Figures 5.5 and 5.6), while 26% are from the Anglosphere (the “en” column in Figures 5.5 and 5.6).

In Figure 5.7 we jointly visualize how “locally” cited each language in our corpus is (x-axis) compared to which portion of citations originate from English-speaking countries (y-axis). Overall, we observe larger variation on the “locality” dimension (values ranging from 0 to 1 with a variance of 0.058) than on the “from English-speaking countries” dimension (values from 0 to 0.67 with a variance of 0.026). Looking at non-Latin script languages, we can see that Cyrillic script languages (e.g., Russian and Ukrainian) are less often of local origin than Asian languages (Chinese, Japanese, Korean) or languages written in Arabic script (Persian¹⁴). Narrowing down on above-mentioned three Asian languages, we observe

¹¹ I.e., documents for which we have MAG metadata (see Table 5.5).

¹² Because a single paper can have authors affiliated with institutions in different locations, we perform our analysis on a per author basis.

¹³ The association between affiliation and country is already given in the MAG. For data on language use per country we refer to the Unicode Common Locale Data Repository's territory-language information (see https://web.archive.org/web/20210225022138/https://unicode-org.github.io/cldr-staging/charts/latest/supplemental/territory_language_information.html [last accessed: 2023-11-10]).

¹⁴ While most varieties of Persian are written in a version of the Arabic script, there also exists varieties written in Cyrillic script [129].

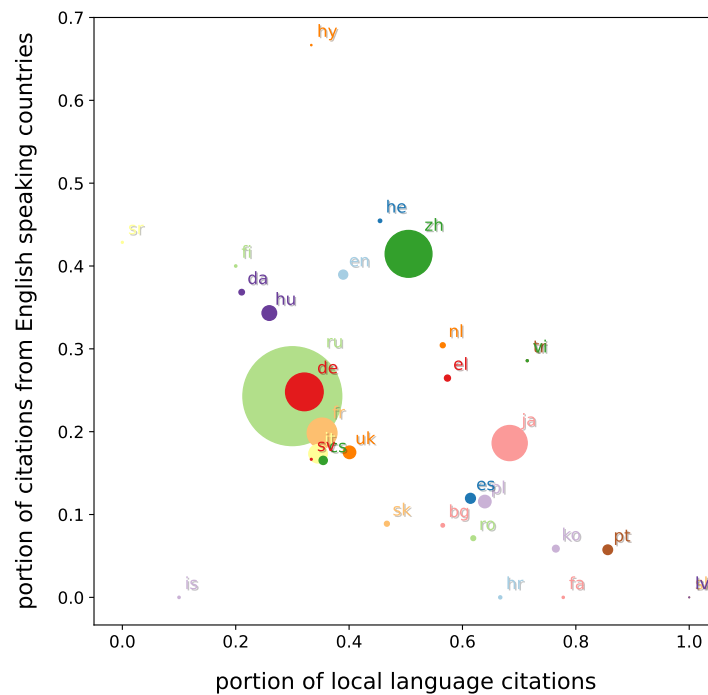


Figure 5.7.: Geographic origin of cross-lingual citations (local vs. English-speaking countries). Marker size (surface area) indicates number of citations.

that for Chinese the relative portion of citations from English-speaking countries (0.41) is more than double of the same measure for Japanese (0.19), which is more than triple the value for Korean (0.06). The comparatively high ratio for Chinese (not just among Asian languages but overall¹⁵) could be taken as an indication for two phenomena: first, an increased relevance of publications written in Chinese (i.e., a higher necessity to cite) and second, an increased rate of scholars able to read Chinese in English-speaking country research institutions (i.e., a higher probability of the ability to cite).

5.5.2.3. Citation Intent and Sentiment

To assess whether or not cross-lingual citations tend to serve a different purpose than their monolingual counterpart, and whether or not authors have a different disposition toward cited works, we analyze the in-text citations (see Figure 5.1) in our corpus.

The analysis of in-text citations—commonly referred to as citation context analysis—is concerned with the textual context of citations [80]. Two tasks in citation context analysis are the classification of citation intent (also referred to as citation function) and citation

¹⁵ The overall comparison has, however, to be done keeping the limitations described in Section 5.4.1 in mind.

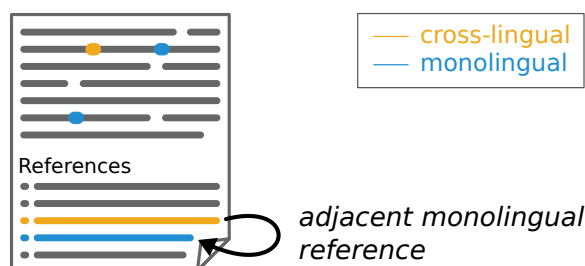


Figure 5.8.: Schematic explanation of an adjacent monolingual reference.

sentiment (also referred to as citation polarity) [80]. Citation intent can reveal why an author added a reference, while the citation sentiment can give insight into the author’s disposition toward that reference. Both citation intent and sentiment have been used in a number of diverse tasks, such as classification [95, 45, 27], summarization [47], and citation recommendation [58]. For citation intent, many schemes have been proposed to classify different functions, ranging from fine-grained to coarse-grained schemes. A partial overview of these can be found in Hernández-Alvarez [80], Jurgens et al. [95], Cohan et al. [45], and Lauscher et al. [113]. These schemes, however, are often domain-specific and too fine-grained [45]. Jurgens et al. [95] proposed a unified scheme of previous work (with six categories), while Cohan et al. [45] proposed a more generalized scheme (with three categories) that works for multiple domains. Recently, Lauscher et al. [113] expanded these schemes to multi-sentence and multi-label citation contexts. Given the number of diverse domains on arXive, we adopt the general scheme by Cohan et al. [45]. For citation sentiment, a three category scheme (*positive*, *negative*, or *neutral*) is widely adopted [20, 14, 80]. Previous approaches to citation intent and sentiment classification have used either hand-crafted rules or classical machine learning models [14, 95], while more recent approaches using deep learning and word embeddings have demonstrated significant improvements in performance [45, 27, 113].

For our analysis, we create two, equally-sized sets of in-text citations. The *in-text x-ling* set (cross-lingual) and the *in-text mono* set (monolingual). In the following we describe the creation of both sets, the classifier model training, and our results for citation intent and sentiment classification.

DATA PREPARATION For the *in-text x-ling* set we determine all in-text citations associated with the references in the cross-lingual set. This yields 45,516 in-text citations for our 33,290 cross-lingual references. The *in-text mono* set is then created by extracting in-text citations associated with adjacent monolingual references. We illustrate this process in Figure 5.8, showing a paper with a single cross-lingual reference for which, accordingly, a single adjacent monolingual reference would be determined and its associated in-text citations (indicated by the two blue markers above) extracted. For *in-text mono* we extract 53,177 in-text citations (i.e., on average more in-text citations per reference) which we reduce to 45,516 through stratified sampling. By sourcing our monolingual in-text citations for comparison from the same papers, we avoid confounding effects such as author specific differences in citation styles.

Table 5.8.: Class distribution and evaluation details for the model training.

Data set	Class	Inst. ^a	Precision	Recall	F1-macro
SciCite	Backgr.	6,375 (58)	86%	93%	86.6%
	Method	3,154 (29)	91%	82%	
	Result	1,491 (13)	86%	83%	
Athar	Neutral	6,901 (87)	91%	98%	67.9%
	Positive	761 (10)	80%	42%	
	Negative	265 (3)	50%	29%	
Athar [†]	Neutral	265 (33)	77%	59%	67.7%
	Positive	265 (33)	59%	59%	
	Negative	265 (33)	65%	94%	
Athar [§]	Neutral	6,901 (90)	96%	97%	82.5%
	Positive	761 (10)	69%	68%	
Athar [‡]	Neutral	761 (50)	85%	69%	80.2%
	Positive	761 (50)	78%	90%	

^a Inst. = Number of instances for training and evaluation (percentage in brackets)

[†] = Under-sampled

[§] = No *Negative* class

[‡] = Under-sampled & no *Negative* class

As a citing sentence can contain more than one citation marker, it is possible that the in-text citations associated with two adjacent reference section entries appear within the same sentence (e.g., as indicated in the second “text” line in Figure 5.8). This is the case for 10,454 of the in-text citations we extracted (i.e., these appear in both sets). We define them as a third set called *mixed*, leaving *in-text x-ling* and *in-text mono* at 35,062 items each.

MODEL TRAINING Training data for citation sentiment and intent classification regarding papers cannot easily be crowdsourced, because domain knowledge is needed for annotation. As a consequence, available data sets are comparatively small. We identify SciCite [45] for citation intent and the data set proposed by Athar [20] for citation sentiment as most appropriate for our purposes.

- SciCite contains 11,020 citations that originate from the Semantic Scholar corpus, which covers several disciplines such as computer science, molecular biology, microbiology and neuroscience [19]. Citations in SciCite are labeled regarding their intent across three categories, namely *Background*, *Method*, and *Result*. The class distribution can be seen in Table 5.8. We select the data set because it is currently the largest available, and classifiers trained on the data set achieve good performance.
- The data set created by Athar contains 8,736 annotated citations from 310 research papers. To the best of our knowledge, it is the largest citation sentiment data set currently available. Following [131], we manually remove 809 items from the data

Table 5.9.: Citation intent and sentiment classification results for cross-lingual, monolingual, and mixed in-text citations. (Values are the number of citations per class followed by the percentage in brackets.)

Data set	Background	Method	Result
<i>x-ling</i>	26,443 (75.4)	7,749 (22.1)	870 (2.5)
<i>mono</i>	26,232 (74.8)	7,801 (22.2)	1,029 (2.9)
<i>mixed</i>	7,688 (73.5)	2,503 (23.9)	263 (2.5)
	Neutral	Positive	Negative
<i>x-ling</i> [*]	34,100 (97.3)	787 (2.2)	175 (0.5)
<i>mono</i> [*]	33,792 (96.4)	1,037 (3.0)	233 (0.7)
<i>mixed</i> [*]	10,049 (96.1)	362 (3.5)	43 (0.4)
<i>x-ling</i> [‡]	22,275 (63.5)	12,787 (36.5)	
<i>mono</i> [‡]	21,825 (62.3)	13,237 (37.8)	
<i>mixed</i> [‡]	6,547 (62.6)	3,907 (37.4)	

^{*} = Classified using the model trained on Athar

[‡] = Classified using the model trained on Athar[‡]

set that are either duplicates or too short to be accurately evaluated regarding their sentiment. The resulting data set, which we refer to as *Athar* from hereon, contains 7,927 citations annotated with one of the three labels *Negative*, *Neutral*, and *Positive*. Citations labeled *Negative* and *Positive* are comparably infrequent in the corpus (see Table 5.8), which makes classifying them more difficult. As possible mitigation strategies, we consider the following options.

- Athar[†]: balancing the data by under-sampling.
- Athar[§]: removing the *Negative* class, as its low performance (see Table 5.8) puts its informativeness into question.
- Athar[‡]: both of the aforementioned.

For each of our classification models, we fine-tune SciBERT [27], a pre-trained language model for scientific text that achieves state-of-the-art performance on sentence classification tasks.

Our evaluation results are shown in Table 5.8. On both SciCite and Athar our models perform on par with the best performing models presented in their respective publications. For citation intent, we achieve an F1 score of 86.6% and relatively similar performance across classes. For citation sentiment, we achieve an F1 score of 67.9% on the original Athar data set. Two of our three class imbalance mitigation strategies (Athar[§] and Athar[‡]) result in an increase in the F1 score to over 80%. Of those two we decide to use the model trained on Athar[‡]. While training on Athar[§] gives us a slightly higher F1 score, the model trained on Athar[‡] achieves high precision and recall for positive citations—which are presumably less common—while also maintaining good performance for neutral citations. Implementation details for the model training can be found in Appendix A.2.

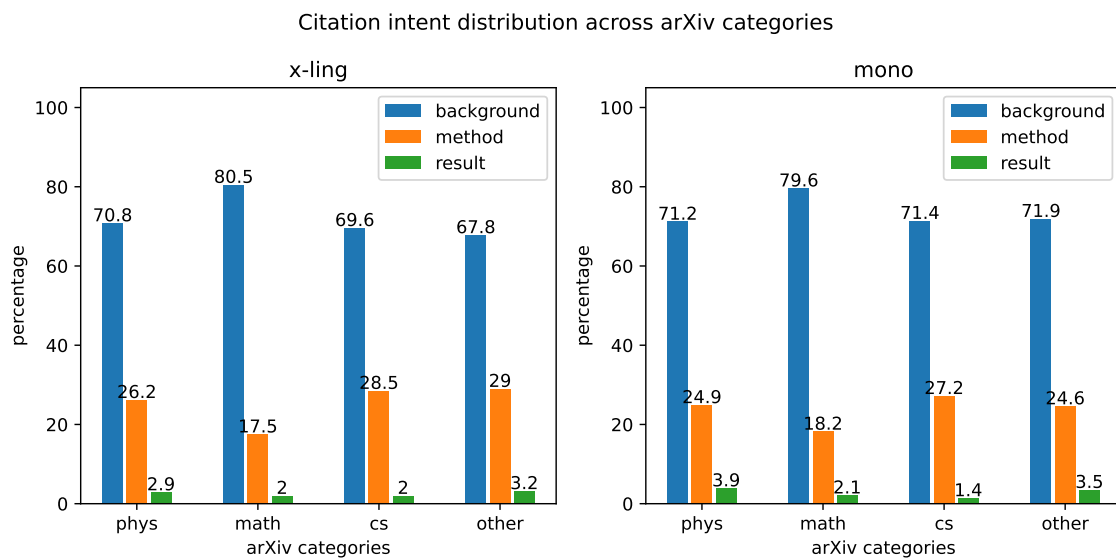


Figure 5.9.: Comparison of citation intent distribution across arXiv categories for *in-text x-ling* (left) and *in-text mono* (right).

CLASSIFICATION RESULTS Based on above evaluation we proceed by using our models trained on SciCite, Athar, and Athar[‡] to classify the intent and sentiment of citations in *in-text x-ling* and *in-text mono*. In Table 5.9, we show the classification results for citation intent (top half) and sentiment (bottom half). The classifiers trained on SciCite and Athar appear to amplify the unbalanced data distribution they were trained on to some degree. Comparing the sentiment classifiers trained on the original Athar and balanced Athar[‡] data set, we see that citations classified as *Positive* increase from around 3% to almost 38%. We take this as a clear sign that reliably distinguishing neutral from positive citations remains a challenge even with state-of-the-art models and training data.

Comparing our results across the data sets *in-text x-ling*, *in-text mono*, and *in-text mixed* we see that in terms of both intent and sentiment class distributions are similar. Taking a closer look at citation intent across the scientific disciplines,¹⁶ we can see in Figure 5.9 that the distributions are overall comparable among disciplines and between cross- and monolingual citations, with mathematics showing a slightly higher use of background citations.

Overall, our results for citation sentiment and intent show no distinct differences between cross- and monolingual citations. This can be taken as an indication for two things. First, that authors cite existing literature with a certain intent and sentiment *regardless* of the cited work’s language. Second, that cross-lingual—while occurring less frequent—serve the same functions as monolingual citations and are therefore not less significant.

¹⁶ We do not evaluate citation sentiment here due to the lacking performance of the sentiment classifiers.

5.5.3. Impact

Regarding the impact of cross-lingual citations, we analyze whether cross-lingual citations in English papers are seen as an “acceptable” practice, whether or not they pose a particular challenge for citation data mining, and their potential impact on the success of the paper they are part of. Our results concerning these three aspects are described in the following sections.

5.5.3.1. Acceptance

To assess the acceptance of cross-lingual citations by the scientific community—that is, whether or not non-English publications are deemed “citable” [117]—we analyze papers in our data that have both a preprint version as well as a published version (in a journal or conference proceedings) dated later than the preprint. This is the case for 2,982 papers. For each preprint-published paper pair, we check if there is a difference in cross-lingual citations. This gives an indication of how the process of peer review affects cross-lingual citations. We perform a manual as well as an automated analysis.¹⁷

For the manual evaluation, we take a random sample of 100 paper pairs. We then retrieve a PDF file of both the preprint and the published version, and manually compare their reference sections. For the automated evaluation, we find that 599 of the 2.9k paper pairs have PDF source URLs given in the MAG. After automatically downloading these and parsing them with GROBID, we are left with 498 valid sets of references. For these, we identify explicitly marked cross-lingual references as described in Section 5.4 and calculate their differences.

Table 5.10 shows the results of our evaluations. In both, cross-lingual citations are more often removed than added, but in the majority of cases left intact. The larger volatility in the automated evaluation is likely due to parsing inconsistencies of GROBID. Our findings complement those of Lillis et al. [117], who, analyzing psychology journals, observe “*some evidence that gatekeepers [...] are explicitly challenging citations in other languages.*” For the fields of physics, mathematics, and computer science, we find no clear indication of a consistent in- or decreasing effect of the peer review process on cross-lingual citations.

5.5.3.2. Impact on Paper Success

To get an indication of whether or not an English paper’s success is influenced by the fact that it contains citations to non-English documents, we compare our cross-lingual set with the random set (see Table 5.4). For both sets we first determine the number of papers that in the MAG metadata have a published version (journal or conference proceedings) in addition to the preprint on arxiv.org. That is, we assume that papers which only have a

¹⁷ Full evaluation details can be found at <https://github.com/Il1Depence/cross-lingual-citations-from-en> [last accessed: 2023-11-10].

Table 5.10.: Changes in cross-lingual citations between preprints and published papers.

Evaluation	#Pairs	#Inc. ^a	#Dec. ^b	Mean ^c	SD ^c
Manual	100	4	7	-0.02	0.529
Automated	498	33	70	-0.12	0.821

^a Inc. = Increased

^b Dec. = Decreased

^c of the differences in the amount of cross-lingual citations

Table 5.11.: Comparison of citations received.

Filter criterion		Cross-lingual set	Random set
-	#Docs	16,300	16,464
	Mean #cit	13.7	18.2
	SD	75.0	51.7
1 ≤ #cit	#Docs	12,074	12,852
and	Mean #cit	12.0	15.1
#cit ≤ 100	SD	15.8	18.4

preprint version did not make it through the peer review process. Using this measure, we observe 9,390 of 16,224 (57.88%) successful papers in the cross-lingual set, and 10,966 of 16,378 (66.96%) successful papers in the random set. Unsurprisingly, due to the higher ratio of published versions, the papers in the random set are also cited more. Table 5.11 shows a comparison of the average number of citations that documents in both sets received. Due to the high standard deviation in the complete sets, we also look at papers which received between 1 and 100 citations, which are comparably frequent in both sets. As we can see, in the unfiltered as well as the filtered case, documents with cross-lingual citations tend to be cited a little less. Because here we can only control for the distribution of papers across years and disciplines, and not for individual authors (as we did in the Section 5.5.2.1), there might be various confounding factors involved.

5.5.3.3. Impact on Citation Data Mining

To assess if cross-lingual citations pose a particular challenge for scholarly data mining—and are therefore likely to be underrepresented in scholarly data—we compare the ratio of references that could be resolved to MAG metadata records for the cross-lingual set and the whole unarXive data set. Of the 39M references in unarXive 42.6% are resolved to a MAG ID. For the complete reference sections of the papers in the cross-lingual set (i.e., references to both non-English and English documents) the number is 45.7% (290,421 of 635,154 references). Looking only at the cross-lingual citations, the success rate of reference resolution drops to 11.2% (3,734 of 33,290 references). We interpret this as a clear

indication that resolving cross-lingual references is a challenge. Possible reasons for this are, for example:

1. A lack of language coverage in the target data set.
For example, if the target data set only contains records of English papers, references to non-English publications cannot be found within and resolved to that target data set.
2. Missing metadata in the target data set.
For example, when there is a primary non-English as well as an alternative English title of a publication, only the former is in the target data set's metadata, but the latter is used in the cross-lingual reference.
3. The use of a title translated "on the fly."
If a non-English publication has no alternative English title, a self translated title in a reference cannot be found in any metadata. To give an example, reference 14 in arXiv:1309.1264 titled "*Hierarchy of reversible logic elements with memory*" is only found in metadata¹⁸ as 記憶付き可逆論理素子の能力の階層構造について.
4. The use of a title transliterated "on the fly."
Similar to an unofficial translated title, if a title is transliterated and this transliteration is not existent in metadata, the provided title is not resolvable. A concrete example of this is the third reference in arXiv:cs/9912004 titled "*Daimeishi-ga Sasumono Sono Sashi-kata*" which is only found in metadata¹⁹ as 代名詞が指すもの,その指し方.

Cases 4 and especially 3 additionally impose a challenge on human readers, as the referred documents can only be found by trying to translate or transliterate back to the original. References to non-English documents which do not have an alternative English title should therefore ideally include enough information to (a) identify the referenced document (i.e., at least the original title), and (b) a way for readers not familiar with the cited document's language to get an idea of what is being cited (e.g., by adding a freely translated English title).²⁰ There are, however, situations where an original title cannot be used. Documents in PubMed Central, for example, cannot contain non-Latin scripts,²¹ meaning that references to documents in Russian, Chinese, Japanese, etc. which do not have alternative English titles are inevitably a challenge for both human readers as well as data mining approaches, unless there is a DOI, URL, or similar identifier that can be referred to.

In light of this, taking a closer look at the 88.8% of unmatched references in the cross-lingual set broken down by languages, we note the following matching failure rates for the five most prevalent languages: Russian: 88.6%, Chinese: 87.0%, Japanese: 91.0%, German:

¹⁸ See <http://hdl.handle.net/2433/172983> [last accessed: 2023-11-10].

¹⁹ See <https://ci.nii.ac.jp/naid/10008827159/> [last accessed: 2023-11-10].

²⁰ An example for this can be found in reference 15 in arXiv:1503.05573: "Шафаревич И. Р. Основы алгебраической геометрии// МЦНМО, Москва, 2007. (English translation: Shafarevich I.R. Foundations of Algebraic Geometry// MCCME, Moscow, 2007)."

²¹ See <https://www.ncbi.nlm.nih.gov/pmc/about/faq/#q16> [last accessed: 2023-11-10].

85.4%, and French: 83.2%. While all of these are high, the numbers for the three non-Latin script languages are noticeably higher than those of German and French. As can be seen with the task of resolving references—and as also indicated through our self-citation data shown in Table 5.7—cross-lingual citations do pose a particular challenge for scholarly data mining.

5.6. Discussion and Conclusion

Utilizing two large data sets, unarXive and the MAG, we performed a large-scale analysis of citations from English papers to non-English language publications (i.e., cross-lingual citations). The data analyzed spans over one million citing publications, 3 disciplines, and 27 years. We gained insights into cross-lingual citations’ prevalence, usage and impact.

Recapitulating our key results, we find that citations to non-Latin script languages can reliably be identified by a “(*in <Language>*)” marker, which enables automated identification in large corpora. Between the disciplines of physics, mathematics, and computer science, cross-lingual citations appear twice as often in mathematics papers compared to the remaining two fields. Over the course of time, we see a downwards trend in citations to Russian and an upwards trend for citations to Chinese. In general, cross-lingual citations are more often of linguistically local origin than originating from English-speaking countries. Citations to Chinese, however, are about twice as likely to come from the Anglosphere than citations to other languages. Concerning authors citing behavior, we observe no remarkable differences between cross- and monolingual citations in terms of self-citations, intent, and sentiment. We also see no clear indication for gatekeeping of cross-lingual citations through the process of peer review. As for the impact of cross-lingual citations on a paper’s success, we only get inconclusive results. Finally, we see clear indicators that cross-lingual citations pose challenges for scholarly data mining, such as a lower likelihood to resolve a cited document due to more complex metadata (e.g., publications having two titles, a primary non-English and an alternative English title) and shortcomings in data integration (e.g., with local citation indices).

Through our preliminary analyses (see Section 5.4.1), we identify challenges in reliably assessing cross-lingual citations to Latin script languages, preventing automated identification in large corpora. These insights can facilitate future efforts in overcoming the identified challenges. Our detailed findings regarding prevalence can help identify scenarios, in which a dedicated effort to take into account cross-lingual citations is warranted. For example, a citation driven analysis of research trends in mathematics might benefit from being able to track “citation trails” into the realm of Russian publications. Lastly, due to the large scale of our investigation, the use of our collected data for machine learning based applications such as cross-lingual citation recommendation is possible.

Our analysis is based on explicit language markers of cited documents, which has shown to be reliable for non-Latin script languages, but only capture a small fraction of citations to Latin script languages. We therefore want to investigate further methods for identifying

cross-lingual citations, to be able to perform more exhaustive analyses. Furthermore, our corpus covers publications from the fields of physics, mathematics, and computer science. While arxiv.org has extensive coverage of physics and mathematics, the share of computer science publications is currently still in a phase of rapid growth. We therefore want to expand our investigation regarding computer science publications to get more representative results, but also include additional disciplines not covered so far. Lastly, we would like to conduct complementary analyses of cross-lingual citations from non-English to English. These might be more challenging to perform on a large scale, because non-English scholarly data is not as readily available. However, such analyses are also likely to yield insights with a larger impact, as citing English language publications is rather common in other languages.

Author Contributions

Tarek Saier: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft (lead), Writing – review & editing. Michael Färber: Supervision, Writing – review & editing. Tornike Tsereteli: Formal analysis, Software, Writing – original draft (support).

Acknowledgements

We thank Irma Suppes for supporting manual data labeling and language identification tasks.


5.7. Result Assessment

The work in this chapter addresses the following research task.

- ❖ **RT3:** *Inclusion of Non-English Publications* - find and implement an approach to include non-English publications into a large-scale, high-quality scholarly data set.

With the presented method to identify cross-lingual citations in English publications, we are able to conduct the so far largest analysis of this type of citation. Our analysis comprises 1.1 M documents and 39 M references. The only analysis comparable in size (0.5 M documents) is restricted to just one non-English language (Russian), and analyses with a similar comparison (citations from English to any language) include fewer than 500 documents and fewer than 20 k references. Through our large-scale identification and analysis of citations into non-English languages, we make significant improvements to counteract Anglocentrism. Accordingly, we deem ❖ **RT3** successfully achieved.

In terms of the overarching research goal of enabling higher-quality scholarly data (see Table 2.4 in Chapter 2), the work presented in this chapter makes the following contributions.

 Scholarly Data Quality Contributions - [4, 5]	
Crit.	Contribution
Rel_{CN}	Language information added to nodes in citation network.
Coy_{CN}	Fine-granular comparison of nodes and edges in citation network enabled by document language information.

Rel_{CN} We determine the language of all cited documents in our corpus. This makes the data as a whole relevant for the study of language related phenomena, such as the ones conducted in this chapter, which are the largest of its kind so far.

Coy_{CN} Providing language data for all cited documents in our corpus makes fine-granular comparison and filtering of nodes and edges in the citation network possible. For example, by applying a respective filter, a subset of only mono- or cross-lingual citations can be created, to be then merged with likewise data.

6

References with Usage Parameters

This chapter is based on the following publication.



Tarek Saier, Mayumi Ohta, Takuto Asakura, and Michael Färber. “HyperPIE: Hyperparameter Information Extraction from Scientific Publications”. In: *Advances in Information Retrieval*. Vol. 14609. Lecture Notes in Computer Science. Springer Nature Switzerland, Mar. 2024, pp. 254–269. ISBN: 978-3-031-56060-6. DOI: 10.1007/978-3-031-56060-6_17

The work in this chapter addresses the following research task.

- ❖ **RT4:** *Fine-grained Research Artifact Representations* - develop a method to extract fine-grained information on research artifacts from text in scientific publications.

6.1. Overview

In this chapter, we present information extraction methods for research artifacts and their usage parameters. This aims to advance the granularity of document representations, as research artifacts become an increasingly relevant object of research. Enabling the extraction of structured information on research artifacts and their parameters enables the study of usage and reporting patterns across time and scientific disciplines. The extracted information furthermore bears potential for use in automated reproduction.

At the end of the chapter, in Section 6.9, we assess the achievement of the research task, as well as the contributions made in terms of the overarching research goal of enabling higher-quality scholarly data.

6.2. Introduction

Models capable of extracting fine-grained information from publications can make scientific knowledge machine-readable at a large scale. Aggregated, such information can

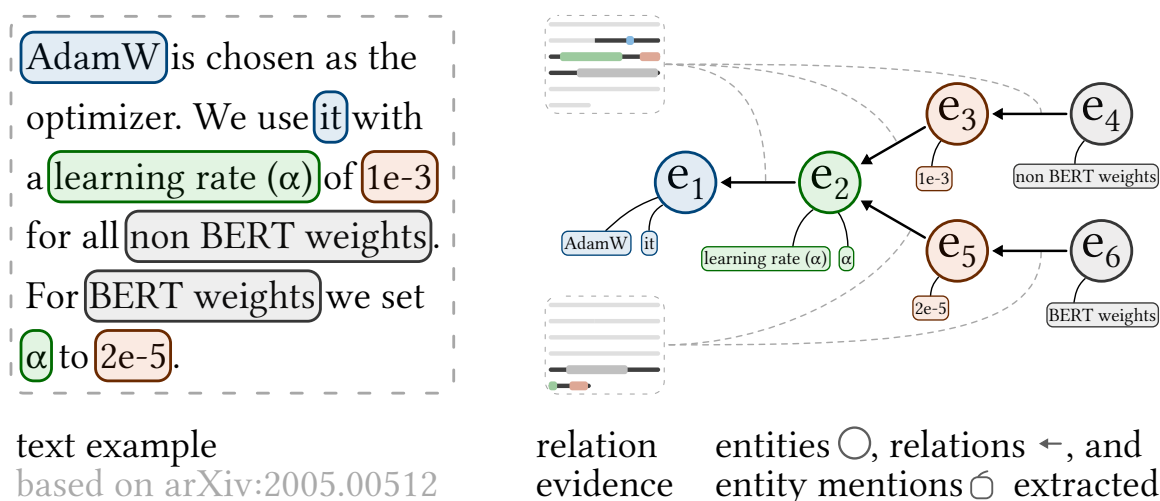


Figure 6.1.: Illustration of hyperparameter information in a text example alongside the extracted entities and relations. Entity types are **research artifact**, **parameter**, **value**, and **context**. Relations are indicated by arrows.

fuel platforms like Papers with Code¹ and the Open Research Knowledge Graph [175, 21], and thereby facilitate academic search, recommendation, and reproducibility. Accordingly, a variety of approaches for information extraction (IE) from scientific text have been proposed [124, 87, 78, 107, 52].

However, to the best of our knowledge, no approaches exist for the extraction of structured information on hyperparameter use from publications. That is, information on *with which parameters* researchers use methods and data. We refer to this information as “hyperparameter information” (see Figure 6.1). Hyperparameter information is important for several reasons. (1) First, its existence in a paper is an indicator for reproducibility [160] and, when extracted automatically, can improve automated reproduction of results [169]. (2) Second, in aggregate it can inform on both conventions in a field as well as trends over time. (3) Lastly, it enables more fine-grained paper representations benefiting downstream applications based on document similarity, such as recommendation and search. Hyperparameter information is challenging to extract, because (1) it is usually reported in a dense format, (2) often includes special notation, and (3) operates on domain specific text (e.g. “For Adam we set α and β to $1e-3$ and 0.9 respectively.”).

To address the lack of approaches for extracting this type of information, we define the task of “hyperparameter information extraction” (HyperPIE) and develop several approaches to it. Specifically, we formalize HyperPIE as an entity recognition (ER) and relation extraction (RE) task. We create a labeled data set spanning a variety of computer science disciplines from machine learning (ML) and related areas. The data set is created by manual annotation of paper full-texts, which is accelerated by a pre-annotation mechanism based on an external knowledge base. Using our data set, we train and evaluate both BERT-based [50] fine-tuned models as well as large language models (LLMs). For the former,

¹ See <https://paperswithcode.com/> [last accessed: 2024-01-10].

we develop a dedicated relation extraction model that achieves an improvement of 29% F_1 compared to a state-of-the-art baseline. For LLMs, we develop an approach leveraging YAML output for structured data extraction, which achieves a consistent improvement in entity recognition across all tested models, averaging at 5.5% F_1 . Using our best performing model, we extract hyperparameter information from 15,000 unannotated papers, and analyze patterns across ML disciplines of how authors report hyperparameters. All our data and source code is made publicly available.² In summary, we make the following contributions.

1. We formalize a novel and relevant IE task (HyperPIE).
2. We create a high-quality, manually labeled data set from paper full-texts, enabling the development and study of approaches to the task.
3. We develop two lines of approaches to HyperPIE and achieve performance improvements in both of them over solutions based on existing work.
4. We demonstrate the utility of our approaches by application on large-scale, unannotated data, and analyze the extracted hyperparameter information.

In the remainder of this chapter we first discuss related work in Section 6.3. We then define the HyperPIE task and describe our data set construction in Section 6.4. This is followed by the description of our BERT- and LLM-based methods in Section 6.5. In Section 6.6 we describe our experiments and results. We conclude with a discussion and overall summary in Sections 6.7 and 6.8, followed by an overarching result assessment in Section 6.9.

6.3. Related Work

6.3.1. Fine-Tuned Models

Named entity recognition (NER) and RE from publications in ML and related fields have been tackled by SciERC [124] and subsequently SciREX [87]. The entity types considered are methods, tasks, data sets, and evaluation metrics. Proposed methods for the task utilize BiLSTMs, BERT and SciBERT [27]. With both approaches, there is a partial overlap in entity types to our task, as we also extract methods and data sets. The key difference arises though the parameter and value entities we cover, which are a challenge in part due to their varied forms of notation (e.g. α / alpha, or $0.001 / 1 \times 10^{-3} / 1e-3$).

IE models aiming to relate natural language to numerical values and mathematical symbols have been introduced at SemEval 2021 Task 8 [78] and SemEval 2022 Task 12 [107] respectively. Most of the proposed models base their processing of natural language on BERT or SciBERT. To handle numbers and symbols rendered in \LaTeX , as well as to accomplish RE between entity types with highly regular writing conventions (e.g. numbers

² See <https://github.com/11lDepence/hyperpie> [last accessed: 2024-01-10].

and units such as “5 ms”), rule-based approaches or dedicated smaller neural networks are commonly used.

Similarly, we find a level of regularity in how authors report parameters and values, and make use of that in our approach accordingly. In line with related work using fine-tuned models, we also use BERT and SciBERT for contextualized token embeddings.

6.3.2. LLMs

With the recent advances in LLMs, there has been a surge in efforts to utilize them for IE from scientific text. Nevertheless, their performance is not on par with dedicated models for NER and RE yet [203].

An important concept for IE with LLMs is introduced by Agrawal et al. [17]: a “resolver” is a function that maps the potentially ambiguous output of an LLM to a defined, task specific output space. In their work, the authors extract singular values and lists from clinical notes using GPT-3. They use a variety of resolvers that perform steps like tokenization, removal of specific symbols or words, and pattern matching using regular expressions.

Work with similar output data complexity (values and lists) has also been done in the area of material science. Xie et al. [200] use GPT-3.5 to extract information on solar cells from paper full-text. Similarly, Polak et al. [153] use ChatGPT to extract material, value, and unit information from sentences of material science papers. They define a conversational progression, in which they prompt the model generate tables, which are processed using simple string parsing rules.

An approach for IE of more complex information is proposed by Dunn et al. [52]. They use GPT-3.5 to extract material information from materials chemistry papers. Given the hierarchical nature of the information to be extracted, the authors find simple output formats insufficient. To overcome this, they prompt the model to output the data in JSON format.³

Given hyperparameter information also is hierarchical (see Figure 6.1), we adopt prompting LLMs to output data in a text based data serialization format. Different from the related work introduced above, we do not limit our experiments to API access based closed source LLMs, but also evaluate various open LLMs, because we recognize the importance of contributing efforts to the advancement of the more transparent, accountable, and reproducibility friendly side of this new and rapidly evolving area of research [116].

Besides IE from scientific publications, there have been efforts to extract hyperparameter schemata and constraints from Python docstrings [25] using CNL grammars [105], and from Python code [161] using static analysis. Compared to our task setting, these rely on a known context (e.g. a `fit` method) and operate on constrained input (generated docstrings and source code instead).

³ See <https://www.json.org/> [last accessed: 2024-01-10].

6.4. Hyperparameter Information Extraction

6.4.1. Task Definition

We define HyperPIE as an ER+RE task with four entity classes “research artifact”, “parameter”, “value”, and “context”, and a single relation type. Briefly illustrated by a minimal example, in the sentence “*During fine-tuning, we use the Adam optimizer with $\alpha = 10^{-4}$.*”, the research artifact *Adam* has the parameter α which is set to the value 10^{-4} in the context *During fine-tuning*.

The entity classes are characterized as follows. A “research artifact”, within the scope of our task, is an entity used for a specific purpose with a set of variable aspects that can be chosen by the user. These include methods, models, and data sets.⁴ A “parameter” is a variable aspect of an artifact. This includes model parameters, but also, for example, the size of a sub-sample of a data set. A “value” expresses a numerical quantity and in our task is treated like an entity rather than a literal. Lastly, a “context” can be attached to a value if the value is only valid in that specific context. The single relation type relates entities as follows: parameter \rightarrow research artifact, value \rightarrow parameter, and context \rightarrow value. Co-reference relations implicitly exist between the mentions of a common entity (e.g. “AdamW” and “it” in Figure 6.1). That is, if an entity has multiple mentions within the text, they are considered co-references to each other.

The scope of the IE task comprises the extraction of entities, their relations, and the identification of all their mentions in the text (and thereby implicitly co-references). Furthermore, we specifically consider IE from text, and not from tables, graphs, or source code.⁵

6.4.2. Data Set Construction

Because HyperPIE is a novel task, we cannot rely on existing data sets for training and evaluating our approaches. We therefore create a new data set by manually annotating papers. As our data source we chose unarXive [3], because it includes paper full-texts and, most importantly, retains mathematical notation as \LaTeX . This is crucial because parsing such notation from PDFs is prone to noise, which would be problematic for our parameter and value entities.

To ensure we cover a wide variety of artifacts and discipline specific writing conventions, we use papers from multiple ML related fields. Specifically, these are Machine Learning

⁴ Broader definitions in other contexts also include software in general, empirical laws, and ideas [118]. For our purposes, however, above specific definition is more useful.

⁵ We leave investigating multi-modal IE pipelines (text/code/graphs) for future work.

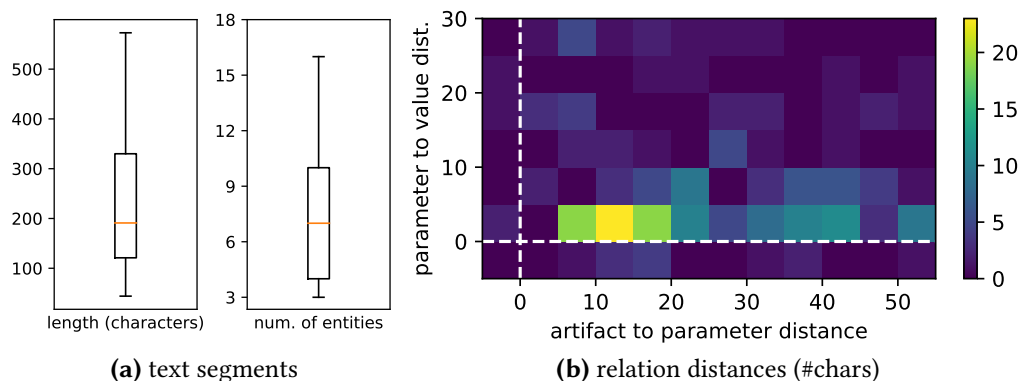


Figure 6.2.: Observations of initial annotation round.

(ML), Computation and Language (CL), Computer Vision (CV), and Digital Libraries (DL), which make up 143,203 papers in unarXive.⁶

We base our annotation guidelines on the widely used ACL RD-TEC guideline⁷ [157]. To make sure our resulting annotations are able to properly capture how authors report hyperparameters in text, we perform two annotation rounds: (1) an initial exploratory round, the results of which are used to refine the annotation guidelines and inform later model development, and (2) the main annotation round, the results of which constitute our data set used for model training and evaluation. In the following, both steps are described in more detail.

6.4.2.1. Initial Annotation Round

We heuristically pre-filter our ML paper corpus for sections reporting on hyperparameters.⁸ Annotators then inspect these sections, select a continuous segment of text that contain hyperparameter information, and make their annotations. This task is performed independently by two annotators and results in a total of 151 text segments (131 unique, 2×10 annotated by both). The annotated text segments contain 1,345 entities and 1,110 relations.

As shown in Figure 6.2a, we observe text segments reporting on hyperparameters to generally have a length below 600 characters. We furthermore see that most text segments contain between 3 and 15 entities. Lastly, in Figure 6.2b, show distances between artifacts and their parameters, as well as parameters and their values. We see that artifacts usually are mentioned before their parameters (78%), and parameters before their values (93%). The reverse cases also exist, but are less common. Additionally, we can see that values are

⁶ The respective arXiv categories are cs.LG, cs.CL, cs.AI, and cs.DL. See https://arxiv.org/category_taxonomy [last accessed: 2024-01-10] for a more detailed description.

⁷ See <https://web.archive.org/web/20220120204209/http://pars.ie/publications/papers/pre-prints/acl-rd-tec-guidelines-ver2.pdf> [last accessed: 2024-01-10].

⁸ We filter based on key phrases (“use”, “set”, etc.), numbers, and \LaTeX math content.

most commonly reported right after their parameter, while there is a higher variability in distances between parameters and artifacts. Based on above observations we determine the unit of annotation for the final round to be one paragraph (on average 563.4 characters long in our corpus), as it is sufficient to capture hyperparameters being reported.

The inter annotator agreement (IAA, reported as Cohen’s kappa) of the text segments annotated by both annotators is 0.867 for entities and 0.737 for relations⁹ (strong to almost perfect agreement) which compares favorably to SciERC [124] with an IAA of 0.769 for entities and 0.678 for relations.

6.4.2.2. Main Annotation Round

In our main annotation round we annotate whole papers (paragraph by paragraph) instead of pre-filtered text-segments. This is done to ensure that the final annotation result reflects data as it will be encountered by a model during inference—that is, containing a realistic amount of paragraphs that have no information on hyperparameters, or, for example, only mention research artifacts but no parameters.

Similar to related work [87], we use Papers with Code as an external knowledge base to pre-annotate entity candidates to make the annotation process more efficient. In a similar fashion, we use annotator’s previously annotated entity mentions for pre-annotation. Pre-annotated text spans are, as the name suggests, set automatically, but need to be checked by annotators manually.

Through this process we annotate 444 paragraphs, which contain 1,971 entities and 614 relations. The entity class distribution is 1,134 research artifacts, 131 parameters, 662 values, and 44 contexts. The annotation data is provided in a JSON structure as shown in Figure 6.1, as well as in the W3C Web Annotation Data Model¹⁰ to facilitate easy re-use and compatibility with existing systems.

6.5. Methods

We approach hyperparameter information extraction in two ways. First, we build upon established ER+RE methods and develop an approach using a fine-tuned model in a supervised learning setting. Second, given the recent promising advances with LLMs, we develop an approach utilizing LLMs in a zero-shot and few-shot setting.

⁹ Measured by the character level entity class and character level relation target span agreement respectively.

¹⁰ See <https://www.w3.org/TR/annotation-model/> [last accessed: 2024-01-10].

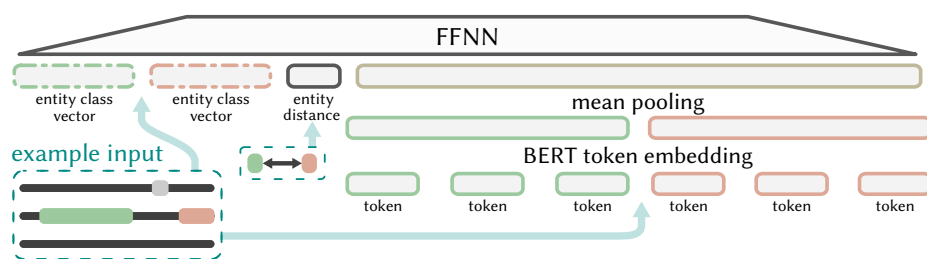


Figure 6.3.: Relation extraction with emphasis on entity candidate pair types and distance.

6.5.1. Fine-Tuned Models

We base our fine-tuned model approach on PL-Marker [204], the currently best performing model on SciERC. Specifically, we use the ER component of PL-Marker. Our reason is that (1) the text our model will be applied on is of the same type as in SciERC (ML publications), and (2) there is some correspondence between the entities to be identified—namely our entity class “research artifact” including methods and datasets, which are both entity classes in SciERC.

For RE we develop an approach that utilizes token embeddings as well as relative entity distance and entity class pairings. This is motivated by the fact that (a) we observed a high level of regularity in the relative distance of research artifact, parameter, and value mentions¹¹ (see Figure 6.2), and (b) relations only exist between specific pairs of entity types.

In Figure 6.3 we show a schematic depiction of our new relation extraction component. Entity candidate pair classes as well as the relative distance between the entities in the text are used as a dedicated model input, BERT token embeddings of the entity mentions are combined using mean pooling. These inputs are fed into a feed-forward neural network FFNN for prediction. Formally, the model performs pairwise binary classification as $\text{FFNN}(E_0^c, E_1^c, E^d, E^T)$, where E_i^c are class vectors, E^d encodes candidate distance, and E^T is the token pair embedding calculated as $E^T = \frac{1}{|T|} \sum_{i=0}^{|T|} \text{BERT}(t_i)$, the mean of the pair’s tokens $t_i \in |T|$.

During the development of our model we also experiment with concatenation in favor of mean pooling to preserve information on the order of the entities, but find that mean pooling results in better performance. Furthermore, we investigated the use of SciBERT instead of BERT, but find that regular BERT embeddings give us better results, despite our model handling scientific text.

¹¹ We note that these observations were made during the initial exploratory annotation round (Section 6.4.2.1) and not during annotation of the evaluation data.

6.5.2. LLMs

We develop our LLM approach for a zero-shot and a few-shot setting. This means the models perform the IE task based on either instructions only (zero-shot), or instructions and a small number of examples (few-shot).

Performing IE using LLMs in zero-shot or few-shot settings requires the desired structure of the output data to be specified within the model input. In simple cases (e.g. numbers or yes/no decisions) this can be achieved by an in-line specification of the format in natural language (e.g. “The answer (Arabic numerals) is”) [101]. IE from scientific publications, however, often seeks to extract more complex information. To achieve this, the model can be tasked to produce output in a text based data serialization format such as JSON, as done in previous work [52]. Especially for complex structured predictions, few-shot prompting has been shown to further boost in-context learning (ICL) accuracy and consistency at inference time [33].

Drawing from techniques used in previous work approaching other IE tasks, we investigate several prompting strategies to build our approach.

1. *Multi-stage prompting* [153]: first determine the presence of hyperparameters information; if present, extract the list of entities; lastly, determine relations.
2. *In-text annotation* [193]: let the input text be repeated with entity annotations, e.g. repeat “We use BERT for ...” as “We use [a1|BERT] for ...”.
3. *Data serialization format* [52]: specify a serialization format in the prompt that is parsed afterwards; then match in-text mentions in the input.
4. (3)+(2): prompt as in (3); then match in-text mentions using (2).

We find (1) to lead to problems with errors propagation along steps. With (2) and (4) we frequently see alterations in the reproduced text. Accordingly, we use prompt type (3) for our approach—specifying a data serialization format in the prompt. While existing work uses the JSON format for this [52], we use YAML, as it is less prone to “delimiter collision” problems due to its minimal requirements for structural characters.¹² In doing so, we expect to avoid problems with LLM output not being parsable. Our overall LLM approach looks as follows.

6.5.2.1. Zero-shot

We build our zero-shot prompts from the following consecutive components: [task] [input text] [format] [completion prefix]. An example is shown in Appendix A.3. In [task] we specify the information to extract, i.e., research artifacts, their parameters, etc. [input text] is the paragraph from which to extract the information. [format] defines the output YAML schema. [completion prefix] is a piece of text that directly

¹² See <https://yaml.org/spec/1.2.2/> [last accessed: 2024-01-10].

precedes the LLM’s output, such as “ASSISTANT: ”. To generate predictions based on LLM output, we pass it to a standard YAML parser after cleansing (e.g. removing text around the YAML block). For each used LLM model, we individually perform prompt tuning. Here we determine, for example, if a model gives better results when the [completion prefix] includes the beginning of the serialized output (e.g. “---\ntext_contains_entities:”) or if this leads to a deterioration in output quality.

6.5.2.2. Few-shot

Our few-shot approach makes the following adjustments to the method described above. Prompts additionally include a component [examples], which are valid input output pairs sampled by their cosine similarity to the input text. Specifically, for an input text from a document X, we sample the five most similar paragraphs from all ground truth documents excluding X. As these examples can be confused with the input text, we reposition the input text to appear *after* the examples. The resulting prompt structure we use for our few-shot approach is as follows: [task] [format] [examples] [input text] [completion prefix]. An example is shown in Appendix A.3.

LLMs reaching a sufficient context size for a few-shot approach to our task are a recent development. We can therefore additionally make use of other recently added capabilities. Specifically, we make use of generation constrains via a gBNF grammar¹³ to enforce LLM output according to our data scheme, allowing us to mitigate parsing errors.

6.6. Experiments

We evaluate the fine-tuned models and LLM approach against baselines from existing work. Both evaluations are performed on our data set described in Section 6.4.2. Metrics used to measure prediction performance are precision, recall and F₁ score, abbreviated as P, R and F₁ respectively.

6.6.1. Fine-Tuned Models

We use PL-Marker, the currently best performing model on SciERC, as our baseline. Models are trained and evaluated using 5-fold cross validation (3 folds training, 1 dev, 1 test). We train the ER component of PL-Marker as done in [204], using *scibert-scivocab-uncased* as the encoder, Adam as the optimizer, a learning rate of 2e-5, and 50 training epochs. Regarding the two RE components we compare, the PL-Marker RE component is trained using *bert-base-uncased*, Adam, a learning rate of 2e-5, and 10 training epochs. Our own RE component also uses *bert-base-uncased*, Adam as the optimizer, and is trained with a

¹³ See <https://github.com/ggerganov/llama.cpp/pull/1773> [last accessed: 2024-01-10].

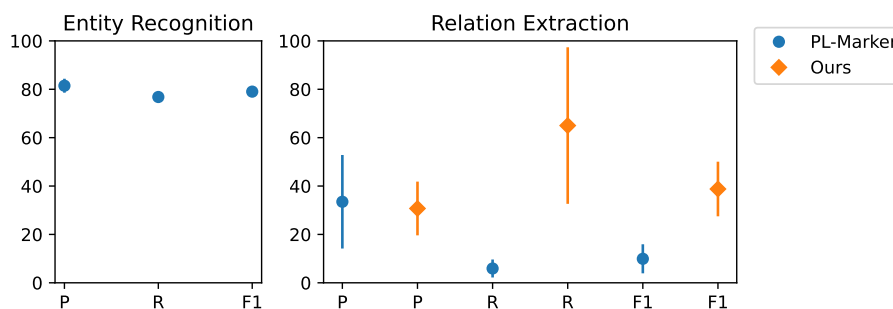


Figure 6.4.: Fine-tuned model evaluation (5-fold cross validation).

Table 6.1.: Ablation study results (model inputs are: T = BERT token embeddings, C = entity class embeddings, D = entity distance).

Model inputs used	P [%]	R [%]	F ₁ [%]
CD	15.5	8.8	11.1
T _D	16.6	29.8	19.6
TC _D	26.5	65.0	35.5
TCD	30.7	65.0	38.8

learning rate of $1e-3$ for 90 epochs.¹⁴ The models are trained and evaluated on a server with a single GeForce RTX 3090 (24 GB).¹⁵

6.6.1.1. Results

In Figure 6.4 we show the results of PL-Marker ER (used for both models) as well as the PL-Marker RE component and our RE model. For ER we evaluate exact matches (no partial token overlap). In the case of RE, each entity pair is predicted as having a relation or not—as there is just one relation type.

Mean ER performance is 81.5, 76.8, and 79.0 (P, R, F₁). For RE, the precision of PL-Marker and our model are similar at 33.5 and 30.7 respectively, but our model performs more consistent. PL-Marker only achieves a very low recall of 5.9, whereas our model, while showing large variability, achieves a mean of 65.0. The resulting F₁ scores are 9.9 for PL-Marker and 38.8 for our model.

6.6.1.2. Analysis

Token level ER performance across entity classes (none, artifact, parameter, value, context) is at 98.5%, 77.8%, 47.9%, 84.4%, 0% F₁. That is, the model does not predict contexts

¹⁴ The two RE models we compare require different learning rates and number of training epochs, because their architecture varies significantly.

¹⁵ More extensive implementation details can be found in Appendix A.3.

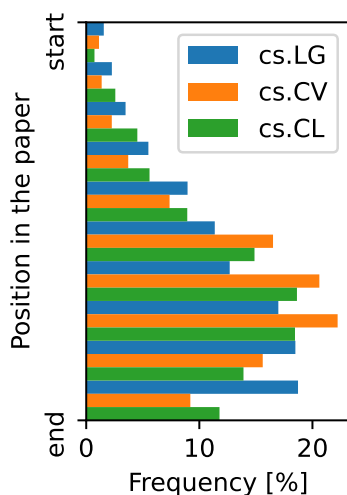


Figure 6.5.: Frequency of hyperparameter mention positions in papers.

Table 6.2.: LLM selection (size in number of parameters).

Model	Variant	Size
WizardLM [201]	WizardLM-13B-V1.1	13 B
Vicuna _{4k} [42]	vicuna-13b-v1.3	13 B
Vicuna _{16k} [42]	vicuna-13b-v1.5-16k	13 B
Falcon [18]	falcon-40b-instruct	40 B
GALACTICA [182]	galactica-120b	120 B
GPT-3.5 [33]	text-davinci-003	175 B

and struggles with parameters, but artifacts and values are predicted reliably. For our RE model, we observe that value-parameter relations are more reliably predicted than parameter-artifact relations.

To assess the impact of the different components in our RE model, we perform an ablation study with the same 5-fold cross-validation setup as above. In Table 6.1, showing its results, we can see that removing the BERT token embeddings (T) results in the largest performance loss, followed by entity class embeddings (C) and entity distance (D). Removing any of the inputs results in worse predictions.

Finally, we apply our full model to a random sample of 15,000 papers. Analyzing the results, we find hyperparameters (artifact, parameter, value triples) are reported in 36% of ML papers, 42% of CV papers, 36% of CL papers, and 7% of DL papers. In Figure 6.5 we further look at the distribution of the information across the length of papers (excluding DL as not being representative). We can see a clear tendency towards the latter half of papers.

Table 6.3.: Prediction performance of LLM models. Subscripts ($\Delta_{\pm n}$) show the delta in F_1 from JSON to YAML output of each model. Format: **best**, second.

Zero-shot		Entity Recognition			Relation Extraction		
Model	Output	P [%]	R [%]	F_1 [%]	P [%]	R [%]	F_1 [%]
WizardLM	JSON	6.9	11.3	8.6	0.1	0.8	0.1
	YAML	9.7	35.6	15.3 $\Delta_{+6.7}$	0.1	1.5	0.1 $\Delta_{+0.0}$
Vicuna _{4k}	JSON	15.1	9.3	11.5	0.7	3.8	1.2
	YAML	17.3	31.5	22.3 $\Delta_{+10.8}$	0.0	0.8	0.1 $\Delta_{-1.1}$
Falcon	JSON	37.1	5.9	10.2	0.0	0.0	0.0
	YAML	32.7	14.2	19.8 $\Delta_{+9.6}$	0.0	0.0	0.0 $\Delta_{+0.0}$
GALACTICA	JSON	25.9	15.7	19.5	0.1	2.3	0.3
	YAML	23.1	19.5	21.1 $\Delta_{+1.6}$	0.0	0.8	0.1 $\Delta_{-0.2}$
GPT-3.5	JSON	27.9	42.8	<u>33.8</u>	<u>5.4</u>	<u>10.7</u>	<u>7.2</u>
	YAML	<u>34.0</u>	<u>41.7</u>	37.4 $\Delta_{+3.6}$	5.8	12.2	7.8 $\Delta_{+0.6}$
5-shot		Entity Recognition			Relation Extraction		
Vicuna _{16k}	JSON	<u>34.4</u>	<u>46.7</u>	<u>39.6</u>	<u>0.8</u>	<u>4.6</u>	<u>1.3</u>
	YAML	43.9	44.1	44.0 $\Delta_{+0.4}$	4.5	9.9	6.1 $\Delta_{+4.8}$

6.6.2. LLMs

For our LLM experiments we chose a variety of models, with sizes ranging from 13 B to 175 B parameters, as shown in Table 6.2. We chose WizardLM [201] as it is meant to handle complex instructions, Vicuna [42] due to its performance relative to its size, Falcon [18] because of its alleged performance, and GALACTICA [182] because it was trained on scientific text. Vicuna_{16k} is a model extended using Position Interpolation [40] based on Rotary Positional Embeddings [178], which makes it the only model in our experiments with a sufficient context size for a few-shot evaluation.

The models are run as follows. GPT-3.5 is accessed through its official API. All open models are run on a high performance compute cluster. Vicuna_{4k} and WizardLM are run on nodes with 4×NVIDIA Tesla V100 (32 GB). GALACTICA, Falcon, and Vicuna_{16k} are run on nodes with 4×NVIDIA A100 (80 GB).¹⁶

As a baseline, we use a JSON variant for each model, where the [format] and [examples] components of prompts use JSON, and compare it to the respective YAML version. All models are used with greedy decoding (temperature = 0) for the sake of reproducibility.

6.6.2.1. Results

In Table 6.3, show the prediction performance of all models and prompt variants. Overall, LLM performance does not reach the level of our pre-trained models. For zero-shot, we observe the best performance with both GPT-3.5 variants, where YAML outperforms

¹⁶ More extensive implementation details can be found in Appendix A.3.

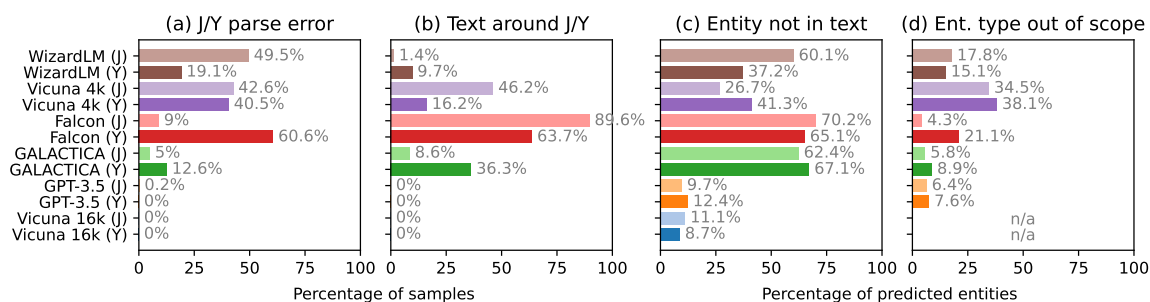


Figure 6.6.: Parsing success, format adherence, hallucinations, and scope adherence of LLM generated JSON (J) and YAML (Y).

JSON (+3.6% ER and +0.6% RE in F_1 score). The second-highest ER F_1 score by model is achieved by Vicuna_{4k} (22.3), despite its size being less than a 10th that of GPT-3.5. For RE, however, even the best model only reaches 7.8%. With our few-shot approach, we are able to considerably improve performance between Vicuna models (+27% ER and +6% RE in F_1), surpassing the zero-shot performance of GPT-3.5 in ER. Lastly, we see that using YAML leads to better ER results across all six models, with ER performance being comparable or improved as well.

6.6.2.2. Analysis

In Figure 6.6 we show an analysis of the steps leading up to model prediction. Focussing first on the zero-shot models (upper five) we observe the following across the four plots from left to right. (a) For three of five models, prompting for YAML leads to fewer parsing errors. (b) Unwanted text around the extracted data is generated more/less by two models each. (c) Hallucinated entities and (d) out of scope entities appear overall slightly more often for in YAML compared to JSON. For our few-shot approach (bottom model), we see that the use of a grammar (a, b) prevents all output format issues. Furthermore (c) hallucinated entities are reduced. (d) Out of scope entities can not be evaluated, because our in-context examples lead to frequent omission of type information in the output.

Through manual analysis we identify a common cause for parsing errors in JSON output to be boolean values (e.g. for “text_contains_entities:”) being copied by the LLM as “true/false” from the prompt. We furthermore find that “entities not in the text” can arise from unsolicited \LaTeX parsing by the LLM (e.g. “\lambda” in text \rightarrow “ λ ” in YAML). Prompting for *verbatim* parameter/value strings does not mitigate this.

6.7. Discussion

Our overall results, with a top performance of 79% F_1 for entity recognition and 39% F_1 for relation extraction, show that extraction of hyperparameter information from scientific text can be accomplished to a degree that yields sound results. There are, however, challenges

that remain, such as more reliable entity recognition of parameters and contexts, as well as more reliable relation extraction in general. Our novel data set enables further development of approaches from hereon. Our IE results on large-scale unannotated data give an indication of possible downstream analyses and applications. Here we see large potential for reproducibility research, faceted search, and recommendation.

Our LLM evaluation shows that for IE tasks dealing with complex information, the choice of text based data serialization format can have a considerable impact on performance, even when using grammar based generation constrains. Additionally, we can see that in-context learning enabled by larger context sizes, as well as grammars, are an effective method to improve IE performance.

Limitations (1) Our work considers HyperPIE from text. This is sensible for a focussed approach, but downstream applications could furthermore benefit from composite pipelines also targeting extraction from tables, source code, etc. (2) We do not test transferability of methods to domains outside of ML related fields. It would require domain expertise to find useful definitions for hyperparameters in each respective domain. (3) Our LLM evaluation does not cover fine-tuning. Presupposing the existence of a large enough training data set, this would be a valuable addition the overall investigation. (4) Defining our YAML/JSON output format hierarchically means that only values associated with parameters and parameters associated with artifacts can be extracted. (5) Lastly, our data and experiments unfortunately are limited to English text only and do not cover other languages.

6.8. Conclusion

We formalize the novel ER+RE task HyperPIE and develop approaches for it, thereby expanding IE from scientific text to hyperparameter information. To this end, we create a manually labeled data set spanning various ML fields. In a supervised learning setting, we propose a BERT-based model that achieves an improvement of 29% F_1 in RE compared to a state-of-the-art baseline. Using the model, we perform IE on a large amount of unannotated papers, and analyze patterns of hyperparameter reporting across ML disciplines. In a zero-/few-shot setting, we propose an LLM based approach using YAML for complex IE, achieving an average improvement of 5.5% F_1 in ER over using JSON. We furthermore achieve large performance gains for LLMs using grammar based generation constrains and in-context learning. In future work, we plan to investigate fine-tuning LLMs, as well as additional practical use cases for data extracted from large publication corpora, such as knowledge graph construction.

Author Contributions

Tarek Saier: Conceptualization, Data curation, Formal analysis, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. Mayumi Ohta: Conceptualization (LLM few-shot), Formal analysis (LLM few-shot), Methodology (LLM few-shot), Software (LLM few-shot), Writing – original draft (support). Takuto Asakura: Conceptualization, Writing – review & editing. Michael Färber: Writing – review & editing.

Acknowledgements

This work was partially supported by the German Federal Ministry of Education and Research (BMBF) via [KOM,BI], a Software Campus project (01IS17042). The authors acknowledge support by the state of Baden-Württemberg through bwHPC. We thank Nicholas Popovic for extensive feedback on the experiment design and prompt engineering. We thank Tarek Gaddour for feedback during the annotation scheme development, and Xiao Ning for input during early model development.

6.9. Result Assessment

The work in this chapter addresses the following research task.

- ❖ **RT4:** *Fine-grained Research Artifact Representations* - develop a method to extract fine-grained information on research artifacts from text in scientific publications.

The presented approaches to extracting hyperparameter information from scientific text achieve an improvement in RE compared to a state-of-the-art baseline (39% compared to 10% in F_1 score). The applicability of the approach on large-scale data is demonstrated through an analysis of hyperparameter reporting patterns in 15,000 unannotated papers. Previous related approaches, namely SciERC and SciREX, only extract comparably shallow information. Specifically, the parameters and parameters' values of research artifacts are not extracted. Our approach adds this information. Accordingly, we deem ❖ **RT4** successfully achieved.

In terms of the overarching research goal of enabling higher-quality scholarly data (see Table 2.4 in Chapter 2), the work presented in this chapter makes the following contributions.

Scholarly Data Quality Contributions - [6]	
Crit.	Contribution
Rel_{SDR}	Added hyperparameter information to the structured document representations
Coy_{SDR}	Fine-granular comparison of documents enabled by through artifacts' parameters and their values

Rel_{SDR} The extracted hyperparameter information from publications' full-text enriches structured document representations with relevant, novel content. Our analysis of hyperparameter reporting patterns in 15,000 papers is a simple demonstration of relevance. Further applications, for which the data is relevant, are, for example, faceted academic search, recommendation, and approaches to automated reproduction.

Coy_{SDR} The added hyperparameter information enables fine-grained comparison of document contents. For example, sections can be compared or filtered based on whether they contain hyperparameter information or not. Furthermore, comparisons based on parameters and their values are made possible.

7

Conclusion

7.1. Summary

This dissertation set out to alleviate the state of scholarly data. To achieve this, the following research objective was set.



Research Objective

Develop an automated process that takes as input scientific publications, and produces as output a high-quality, machine-readable derivative representation of the publications.

Criteria for high quality in the context of scholarly data were defined across the following five dimensions.



Data Quality Dimensions

(1) relevance, (2) accuracy, (3) timeliness, (4) comparability, (5) completeness

For each dimension, criteria were defined for scholarly data's key elements: the citation network (CN) and structured document representations (SDR), as shown in Table 7.1.

Table 7.1.: Scholarly Data Quality Criteria.

Dimension	Focus	Specific Criterion and Short Description
Relevance	CN	Rel_{CN} Representative coverage of publications in area of study
	SDR	Rel_{SDR} Inclusion of relevant content types (text, math, etc.)
Accuracy	CN	Acc_{CN} Correctly linked references
	SDR	Acc_{SDR} Noise-free full-text content
Timeliness	both	Tim_{C/S} Coverage of recent publications
Comparability	CN	Co_{CN} Use of established doc. identifiers (DOI, PMID, etc.)
	SDR	Co_{SDR} Fine-granular, specifically typed content representation
Completeness	CN	Co_{CN} All references in publications successfully linked
	SDR	Co_{SDR} No sections or content missing (appendices, math, etc.)

To focus the efforts of improving scholarly data quality, we identified three areas of key importance, in which current scholarly data has significant limitations (see Section 1.4).

1. **Citation Network**

Lacking completeness of the network connecting publications through citations.

2. **Anglocentrism**

Lacking coverage of non-English publications in data used and analyzed.

3. **Research Artifacts**

Lack of structured representation of research artifacts mentioned in publications.

In order to address these, four research tasks were defined.

- ❖ **RT1:** *Base Methodology* - establish a base methodology for generating a large-scale, high-quality scholarly data set, that is on par with or improving upon existing data sets.
- ❖ **RT2:** *Citation Network Completeness* - develop a method to link literature references, that is able to link more references than are linked in existing data sets, while not compromising on link correctness or processing efficiency.
- ❖ **RT3:** *Inclusion of Non-English Publications* - find and implement an approach to include non-English publications into a large-scale, high-quality scholarly data set.
- ❖ **RT4:** *Fine-grained Research Artifact Representations* - develop a method to extract fine-grained information on research artifacts from text in scientific publications.

The four research tasks were accomplished as follows. In Chapter 3, we developed a corpus creation method transforming publications' \LaTeX source files into a large-scale corpus of interlinked, annotated, full-text documents ❖ **RT1**✓. The presented method also includes a highly accurate reference matching procedure ❖ **RT2**(✓). Applying our method on the complete set of all publications on arXiv.org, we created the data set *unarXive*, which was used as a basis for all subsequent work. In Chapter 4, we presented improvements regarding the citation network and the granularity of document representations. For the citation network, a blocking technique was developed that, applied on the set of references in a corpus, increases the number of matched references and bibliographic couplings. With an updated corpus creation method, the *unarXive* data set achieved a more complete, state-of-the-art citation network ❖ **RT2**✓. The updated procedure furthermore enabled more fine-grained document representations, which in turn made the subsequent work in Chapter 6 possible. In Chapter 5, we studied cross-lingual citations in the *unarXive* corpus. For this, we developed a method to reliably identify this type of citation based on raw reference strings. In our study, which is the largest of its kind to date, we analyzed cross-lingual citations' prevalence, usage, and impact ❖ **RT3**✓. Lastly, in Chapter 6, we developed methods for extracting information about research artifacts and their usage parameters from publication full-texts. Applying our best performing method on *unarXive*, we found differences in parameter reporting patterns across several disciplines ❖ **RT4**✓.

Through the accomplishment of the four research tasks, significant improvements across all quality dimensions and criteria were achieved, as summarized in the overview below.

Scholarly Data Quality Contributions - Overview						
Quality Dimension	Criterion	Contribution				
Relevance	Rel_{CN}	+	=	+	○	
	Rel_{SDR}	○	+	○	+	
Accuracy	Acc_{CN}	+	=	○	○	
	Acc_{SDR}	+	=	○	○	
Timeliness	Tim_{C/S}	+	+	○	○	
Comparability	Coy_{CN}	+	+	+	○	
	Coy_{SDR}	○	+	○	+	
Completeness	Cos_{CN}	+	+	○	○	
	Cos_{SDR}	+	+	○	○	
		Chapter	3	4	5	6
		Publication	[1]	[2, 3]	[4, 5]	[6]

Legend
 +: SOTA/improvement/etc. (see respective chapter)
 =: equal to previous
 ○: not considered

In summary, we successfully addressed three key areas of limitation of current scholarly data and achieved comprehensive improvements in data quality. All introduced methods operate automatically and are demonstrably applicable to large-scale data. Accordingly, we argue to have succeeded in developing an automated process producing high-quality derivative representation of scientific publications, and therefore to have accomplished our **Research Objective** ✓.

Naturally, advances uncover new challenges, and room for improvement remains. In the following section, we therefore discuss the impact and limitations of our work, as well as prospective avenues for further improvement.

7.2. Discussion

We discuss the contributions made in this dissertation from three different perspectives: (1) the identified research gap, (2) the five quality dimensions, and (3) the research community.

7.2.1. Research Gaps

We made significant progress in all three areas of the identified research gap.

1. Citation Network

Achieving state-of-the-art citation network completeness on a large-scale, multi-discipline corpus, means we enable analysis results more valid than previously possible, and the training of prediction models grounded more in reality than before. That being said, our achieved completeness of 44.4% means that missing citation links in scholarly data remain a problem. We see potential for future improvements in the development of sophisticated inter-reference blocking and matching methods, building upon our work in Chapter 4.

2. Anglocentrism

Performing the largest study on citations of non-English publications to date, we provide novel insight into an understudied phenomenon, and are able to highlight challenges for the integration of scholarly data across language borders. Based on our findings, we see potential for improvements through better language support of platforms, and more widespread use of unique document identifiers. Another important aspect is the development of information extraction approaches applicable to non-English publications, which our work does not cover.

3. Research Artifacts

With our task definition as well as model development and application for the extraction of hyperparameter information, we enable the use and study of an important type of content in scientific publications. Our model performance of 79% F_1 for entity recognition and 39% for relation extraction indicates remaining challenges particular regarding the latter. Based on our analysis, we can point to parameter entities as a particularly viable focus for achieving improvements in future endeavors.

7.2.2. Quality Dimensions

Through addressing the identified research gap, our work achieves comprehensive improvements of data quality, as determined across five dimensions.

1. Relevance

Our improvements regarding relevance stem from two areas. (i) We cover a large extent of documents with clear assignability to a subject of study, thereby facilitating, for example, representative coverage of an area of research. This improvement is by virtue of making use of arXiv as a data source. (ii) We furthermore enable the extraction and structured representation of significant document contents, such as hyperparameter information. This improvement is independent of the specific data source used.

2. Accuracy

Improvements we achieve in document representation accuracy are accomplished by harnessing a partially structured data source. In particular, we perform our work based on papers' \LaTeX sources. However, similar source types such as JATS XML and DOCX files bear the same potential. The citation network accuracy of >96% that we achieve is already very high, with identified errors being edge cases such as follow-up publications with near identical titles.

3. Timeliness

Our improvements in terms of data timeliness are a result of us updating our corpus to include recent publications (e.g. up until the end of the most recent completed year). This level of timeliness is arguably sufficient for the study of and applications based on phenomena that do not change within the span of a year, such as citing behavior or writing conventions (e.g. how hyperparameters are reported). Furthermore, a data set with a fixed set of contents is beneficial for comparison of approaches on the same data. For some applications, however, it is desirable to have data on publications included right with their release. An example for this would be paper recommendation. In such cases a "living corpus" that is constantly updated is preferable.

4. Comparability

We achieve improvements in comparability primarily based on (i) determining documents' unique identifiers, and (ii) providing fine-granular structure in our document content representations. (i) Regarding document identifiers, DOIs are most established in academia, but a significant portion of publications without DOI exists (e.g. measured in 2014 on Web of Science and Scopus at 12% in life sciences, 15% in physical & health sciences, and 23% in social sciences & humanities [71]). By providing additional identifiers, we can cover part of those as well. (ii) As for document content, the typed section and paragraph structure we provide in *unarXive* represents natural semantic units on the intra-document level. On the level of sentences, a wide range of structures of interest can be conceived of. Our choice to focus on hyperparameter information is motivated by considerations of potential impact.

5. Completeness

Improvements we achieve in terms of data completeness stem from our reference matching—the results of which we already discussed in Section 7.2.1 above—, and our \LaTeX document conversion methodology. Regarding the latter, we are able to provide some document content in addition to the full-text—namely captions of tables and figures as well as mathematical notation. Tables and figures themselves would be a valuable addition, but require the development of additional extraction mechanisms and were not considered.

7.2.3. Research Community

Despite its recency, our work already made an impact on the research fields concerned with scholarly data and the study of publications. Below, we give a brief account of ideas and results from this dissertation permeating into and being used by the research community.

- Use of **methodology**
 - In [122] Lo et al. use our corpus creation methodology for creating the \LaTeX subset of their S2ORC data set.
 - Chen et al. build on our reference matching procedure in [38] for the creation of their SciXGen data set.
- Use for **model development and evaluation**
 - Meyer et al. use our data for the development and evaluation of a citation recommendation model in [132].
 - In [151], Parisot and Zavrel train a novel multi-objective representation learning technique for scientific document retrieval on our data.
 - With Researcher2Vec, Mochihashi presents a method for researcher profile embeddings in [136], using our data to validate their approach.
- Use for **analyses**
 - In [188] Veneri et al. use our data to investigate how astronomers cite other research fields.
 - Xue uses our data to analyze in [202] semantic shifts of the contexts in which works are cited.
 - Meng et al. use our data in [130] for an analysis of omitted citations of works that have become common knowledge – so called “obliteration by incorporation”.

Comparing our work to existing efforts within the research community which strive to create high-quality scholarly data, we find that our work particularly stands out through the *combination* of the following three aspects. (1) Accurate, fine-granular document representations, (2) a citation network, and (3) applicability on a large scale due to being automated. This distinguishes our work from existing efforts as follows. S2ORC [122] predominantly uses PDF data and therefore does not provide the same level of granularity (e.g. mathematical notation) and is more prone to noise. arXMLiv [68], while providing accurate, fine-granular document representations, lacks a citation network. Lastly, the Open Research Knowledge Graph (ORKG) [175, 21] relies on manual or only semi-automated adding of data, and is therefore limited in scale.

Our work, as well as the related work described above, seek to represent scientific publications in a broad way. That is, multiple scientific disciplines and large time spans are covered, and the structured data reflects multiple aspects such as full-text, citation network,

authors, etc. Another approach towards high-quality scholarly data are dedicated efforts in specific areas. An example of such an effort is the OpenCitations Index,¹ focussing solely on citation data. Such an approach, however, necessitates the ability to combine multiple dedicated resources. For example, combining citation data with publication full-texts. This is only possible as far as unique persistent identifiers for all involved entities exist—e.g. DOIs for documents, ORCiDs for authors, and ROR IDs for affiliations [128]. Although use of such identifiers is becoming more and more established [71], gaps in their coverage mean that, for the moment, a combination of dedicated resources is only of limited use [205, 75].

To briefly recap, we discussed our contributions and impact (1) in terms of the addressed research gap, (2) across the five data quality dimensions, and (3) in relation to the research community. Overall, our work constitutes a range of measurable and demonstrated advancements, and has furthermore been taken up by the research community.

7.3. Outlook

We conclude with a brief look at viable extensions of our work, as well as potential future developments in scientific publishing and what they would mean for the presented work.

7.3.1. Extensions of Our Work

Extension to Other Input Formats Our work takes publications' \LaTeX source files from arXiv.org as the starting point. Because \LaTeX provides a certain level of explicit document structure and semantic information, a natural extension would be to replicate our work on the likewise structured JATS XML publications of the PubMed Central Open Access Subset. This would widen the scope of the results attained, namely by adding life sciences to the already covered disciplines of physics, mathematics, and computer science. In disciplines other than the aforementioned, however, only PDFs are available in large quantities, given the current state of scientific publishing. An extension to PDF input would likely come with challenges regarding structured document representations. However, the work we presented for the focus areas *citation network*, *non-English content*, and *research artifacts*—see Chapters 4, 5 and 6 respectively—is not reliant on \LaTeX as a starting point. This means, given methods for information extraction from PDFs producing output equal to our intermediate results from \LaTeX , extending our work to PDF based document collections is likely possible without major challenges.

¹ See <https://opencitations.net/index> [last accessed: 2023-11-30].

Reference Parsing and Matching We achieve state-of-the-art citation network completeness, but the amount of missing citation links in scholarly data remains an issue. Regarding future development, we see three key elements playing a role. First, the parsing of reference strings in order to extract structured information (title, authors, etc.) bears potential for improvement using synthetic training data [10]. Second, based on our work on reference clustering presented in Chapter 4, the development and application of novel clustering approaches is promising. Third, a continuation of the trend that DOIs usage is becoming more and more established [71] can be expected to simplify the underlying challenge itself, at least regarding future publications.

Integration of Non-English Publication Repositories We studied references *to non-English publications* in the English full-text documents in our corpus. More extensive follow-up studies could be made possible by integrating large repositories of non-English publications, such as the Japanese J-STAGE² containing over 5 million open access articles. This would enable, for example, studying the “reverse” phenomenon of what we examined, and analyze references *from non-English* publications. Furthermore, the resulting large-scale multilingual full-text corpus would be a valuable resource for the development and evaluation of information extraction models not limited to English, thereby further counteracting Anglocentrism in scholarly data related research.

Utilization of Hyperparameter Information We developed models for the extraction of hyperparameter information from papers’ full-text. To demonstrate their applicability on large-scale data, we perform an exemplary analysis of differences in hyperparameter reporting patterns across disciplines. Beyond this exemplary use, the extracted information bears potential for powering faceted academic search and recommendation systems, as well as the development of approaches to automated reproduction. However, because the performance of our models is still limited, especially in terms of relations extraction when parameter type entities are involved, we see further efforts towards model improvement as a priority.

7.3.2. Future External Developments

LLMs Regarding information extraction methodologies, a continuation of the recent advances in LLM technology could become a key factor in bridging the gap between our by-human for-human publications, and machine-readable scholarly data. This is because, even though LLM performance is not on par with dedicated models yet [203], the wide-ranging information available to them could allow filling in the assumed background knowledge that is necessary for understanding, but not explicitly mentioned in, scientific publications. However, a particular challenge with the use of LLMs for the creation of

² See <https://www.jstage.jst.go.jp/> [last accessed: 2024-02-07].

scholarly data, is scaling approaches to large-scale data, because of the required computing resources.

Tagged PDFs Looking beyond the current state of scholarly data, where it is necessary to apply information extraction methods to retroactively determine document structure and semantic information, future developments concerning “tagged PDFs” [114] could simplify the creation of high-quality scholarly data. Widespread adoption of encouragement or requirements for semantically tagging PDFs could either be driven by efforts to improve the accessibility of scientific publications, or by the fact that it would facilitate data mining. In STEM fields, a prerequisite for such a development would be that the L^AT_EX Project’s plan to support semantic annotation natively succeeds [135, 134].

What Constitutes a Publication Considering future developments of the landscape of scientific publications, changes to the status quo of papers being the primary unit of publication would accordingly bring changes to the nature of scholarly data. For example, establishment of micropublications [158] and further adoption of data citations [103] could bring new requirements and opportunities to scholarly data, both in terms of data modeling as well as information extraction methods.

Through the continuation of our work and that of our colleagues, we envision a gradual closing of the gap between scientific publications and their machine-readable representation in the form of scholarly data, eventually enabling a digital record of science faithful to its anthropocentric origin. This dissertation marks a substantial step along this path.

Bibliography of Publications

- [1] **Tarek Saier** and Michael Färber. “unarXive: A Large Scholarly Data Set with Publications’ Full-Text, Annotated In-Text Citations, and Links to Metadata”. In: *Scientometrics* 125.3 (Dec. 2020), pp. 3085–3108. ISSN: 1588-2861. DOI: 10.1007/s11192-020-03382-z.
- [2] **Tarek Saier**, Meng Luan, and Michael Färber. “A Blocking-Based Approach to Enhance Large-Scale Reference Linking”. In: *Proceedings of the workshop on understanding literature references in academic full text (ULITE) at JCDL 2022*. June 2022.
- [3] **Tarek Saier**, Johan Krause, and Michael Färber. “unarXive 2022: All arXiv Publications Pre-Processed for NLP, Including Structured Full-Text and Citation Network”. In: *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE Computer Society, June 2023, pp. 66–70. DOI: 10.1109/JCDL57899.2023.00020.
- [4] **Tarek Saier** and Michael Färber. “A Large-Scale Analysis of Cross-lingual Citations in English Papers”. In: *Digital Libraries at Times of Massive Societal Transition*. Springer International Publishing, 2020, pp. 122–138. ISBN: 978-3-030-64452-9. DOI: 10.1007/978-3-030-64452-9_11.
- [5] **Tarek Saier**, Michael Färber, and Tornike Tsereteli. “Cross-Lingual Citations in English Papers: A Large-Scale Analysis of Prevalence, Usage, and Impact”. In: *International Journal on Digital Libraries* 23.2 (June 2022), pp. 179–195. ISSN: 1432-1300. DOI: 10.1007/s00799-021-00312-z.
- [6] **Tarek Saier**, Mayumi Ohta, Takuto Asakura, and Michael Färber. “HyperPIE: Hyperparameter Information Extraction from Scientific Publications”. In: *Advances in Information Retrieval*. Vol. 14609. Lecture Notes in Computer Science. Springer Nature Switzerland, Mar. 2024, pp. 254–269. ISBN: 978-3-031-56060-6. DOI: 10.1007/978-3-031-56060-6_17.
- [7] **Tarek Saier** and Michael Färber. “Bibliometric-Enhanced arXiv: A Data Set for Paper-Based and Citation-Based Tasks”. In: *Proceedings of the 8th International Workshop on Bibliometric-enhanced Information Retrieval (BIR) co-located with the 41st European Conference on Information Retrieval (ECIR 2019)*. BIR’19. Cologne, Germany, 2019, pp. 14–26.
- [8] **Tarek Saier** and Michael Färber. “Semantic Modelling of Citation Contexts for Context-Aware Citation Recommendation”. In: *Advances in Information Retrieval*. Springer International Publishing, 2020, pp. 220–233. DOI: 10.1007/978-3-030-45439-5_15.
- [9] Johan Krause, Igor Shapiro, **Tarek Saier**, and Michael Färber. “Bootstrapping Multilingual Metadata Extraction: A Showcase in Cyrillic”. In: *Proceedings of the Second Workshop on Scholarly Document Processing*. Online: Association for Computational Linguistics, June 2021, pp. 66–72. DOI: 10.18653/v1/2021.sdp-1.8. URL: <https://aclanthology.org/2021.sdp-1.8>.

- [10] Igor Shapiro, **Tarek Saier**, and Michael Färber. “Sequence Labeling for Citation Field Extraction from Cyrillic Script References”. In: *SDU 2022: Scientific Document Understanding 2022; Proceedings of the Workshop on Scientific Document Understanding; co-located with 36th AAAI Conference on Artificial Intelligence (AAAI 2022)*. Mar. 2022.
- [11] Michael Färber, Christoph Braun, Nicholas Popovic, **Tarek Saier**, and Kristian Noullet. “Which Publications’ Metadata Are in Which Bibliographic Databases? A System for Exploration”. In: *Bibliometric-enhanced Information Retrieval, Proceedings of the 12th International Workshop on Bibliometric-enhanced Information Retrieval co-located with 44th European Conference on Information Retrieval (ECIR 2022)*. Apr. 2022, pp. 39–44.
- [12] Chifumi Nishioka, Michael Färber, and **Tarek Saier**. “How Does Author Affiliation Affect Preprint Citation Count? Analyzing Citation Bias at the Institution and Country Level”. In: *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries. JCDL ’22*. Association for Computing Machinery, 2022. ISBN: 9781450393454. DOI: 10.1145/3529372.3530953.
- [13] **Tarek Saier**, Youxiang Dong, and Michael Färber. “CoCon: A Data Set on Combined Contextualized Research Artifact Use”. In: *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. Los Alamitos, CA, USA: IEEE Computer Society, June 2023, pp. 47–50. DOI: 10.1109/JCDL57899.2023.00016.

Bibliography of References

- [14] Amjad Abu-Jbara, Jefferson Ezra, and Dragomir Radev. “Purpose and Polarity of Citation: Towards NLP-based Bibliometrics”. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, 2013, pp. 596–606.
- [15] Amjad Abu-Jbara and Dragomir Radev. “Reference Scope Identification in Citing Sentences”. In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Montreal, Canada: Association for Computational Linguistics, 2012, pp. 80–90. ISBN: 978-1-937284-20-6.
- [16] Charu C. Aggarwal. *Machine Learning for Text*. Springer International Publishing, Apr. 2018. ISBN: 978-3-319-73530-6. DOI: 10.1007/978-3-319-73531-3.
- [17] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. “Large language models are few-shot clinical information extractors”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Dec. 2022, pp. 1998–2022.
- [18] Ebtesam Almazrouei et al. *Falcon-40B: an open large language model with state-of-the-art performance*. 2023.
- [19] Waleed Ammar et al. “Construction of the Literature Graph in Semantic Scholar”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*. Association for Computational Linguistics, June 2018, pp. 84–91. DOI: 10.18653/v1/N18-3011. URL: <https://www.aclweb.org/anthology/N18-3011>.
- [20] Awais Athar. “Sentiment Analysis of Citations using Sentence Structure-Based Features”. In: *Proceedings of the ACL 2011 Student Session*. Portland, OR, USA: Association for Computational Linguistics, June 2011, pp. 81–87. URL: <https://www.aclweb.org/anthology/P11-3015>.
- [21] Sören Auer et al. “Improving Access to Scientific Literature with Knowledge Graphs”. In: *Bibliothek Forschung und Praxis* 44.3 (2020), pp. 516–529. DOI: 10.1515/bfp-2020-2042.
- [22] Chhandak Bagchi, Eric Malmi, and Przemyslaw Grabowicz. *Promotion of Scientific Publications on ArXiv and X Is on the Rise and Impacts Citations*. Jan. 2024. DOI: 10.48550/arXiv.2401.11116.

- [23] Elisa Bassignana and Barbara Plank. “What Do You Mean by Relation Extraction? A Survey on Datasets and Study on Scientific Relation Classification”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Ed. by Samuel Louvan, Andrea Madotto, and Brielen Madureira. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 67–83. DOI: 10.18653/v1/2022.acl-srw.7.
- [24] Hannah Bast and Claudius Korzen. “A Benchmark and Evaluation for Text Extraction from PDF”. In: *Proceedings of the 2017 ACM/IEEE Joint Conference on Digital Libraries*. JCDL’17. Toronto, ON, Canada, 2017, pp. 99–108. DOI: 10.1109/JCDL.2017.7991564.
- [25] Guillaume Baudart, Peter D. Kirchner, Martin Hirzel, and Kiran Kate. “Mining Documentation to Extract Hyperparameter Schemas”. In: *Proceedings of the 7th ICML Workshop on Automated Machine Learning (AutoML 2020)*. 2020.
- [26] Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. “Research-paper recommender systems: a literature survey”. In: *International Journal on Digital Libraries* 17.4 (Nov. 2016), pp. 305–338. ISSN: 1432-1300. DOI: 10.1007/s00799-015-0156-0.
- [27] Iz Beltagy, Kyle Lo, and Arman Cohan. “SciBERT: A Pretrained Language Model for Scientific Text”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3615–3620. DOI: 10.18653/v1/D19-1371. URL: <https://www.aclweb.org/anthology/D19-1371>.
- [28] Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. “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*. Association for Computational Linguistics, June 2018, pp. 238–251. DOI: 10.18653/v1/N18-1022.
- [29] Steven Bird et al. “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, 2008.
- [30] Christof Bless, Ildar Baimuratov, and Oliver Karras. “SciKGT_{EX} - A LATEX Package to Semantically Annotate Contributions in Scientific Publications”. In: *ACM/IEEE Joint Conference on Digital Libraries, JCDL 2023*. IEEE, June 2023, pp. 155–164. DOI: 10.1109/JCDL57899.2023.00030.
- [31] Kevin W. Boyack and Richard Klavans. “Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?” In: *Journal of the American Society for Information Science and Technology* 61.12 (2010), pp. 2389–2404. ISSN: 1532-2890. DOI: 10.1002/asi.21419.
- [32] Lawrence D. Brown, T. Tony Cai, and Anirban DasGupta. “Interval Estimation for a Binomial Proportion”. In: *Statistical Science* 16.2 (2001), pp. 101–133.

- [33] Tom B. Brown et al. “Language Models Are Few-Shot Learners”. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS’20. 2020.
- [34] Cornelia Caragea et al. “CiteSeer x : A Scholarly Big Dataset”. In: *Proceedings of the 36th European Conference on IR Research*. ECIR’14. Amsterdam, The Netherlands, 2014, pp. 311–322.
- [35] Tanmoy Chakraborty and Ramasuri Narayanam. “All Fingers are not Equal: Intensity of References in Scientific Articles”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. EMNLP’16. Austin, Texas, USA, 2016, pp. 1348–1358.
- [36] Muthu Kumar Chandrasekaran, Michihiro Yasunaga, Dragomir R. Radev, Dayne Freitag, and Min-Yen Kan. “Overview and Results: CL-SciSumm Shared Task 2019”. In: *Proceedings of the 4th Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries*. BIRNDL’19. Paris, France, 2019, pp. 153–166.
- [37] Chaomei Chen. “CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature”. In: *Journal of the American Society for Information Science and Technology* 57.3 (2006), pp. 359–377. DOI: 10.1002/asi.20317.
- [38] Hong Chen, Hiroya Takamura, and Hideki Nakayama. “SciXGen: A Scientific Paper Dataset for Context-Aware Text Generation”. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Nov. 2021, pp. 1483–1492. DOI: 10.18653/v1/2021.findings-emnlp.128.
- [39] Jingqiang Chen and Hai Zhuge. “Automatic generation of related work through summarizing citations”. In: *Concurrency and Computation: Practice and Experience* 31.3 (2019).
- [40] Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. “Extending context window of large language models via positional interpolation”. In: *arXiv preprint arXiv:2306.15595* (2023).
- [41] Pei-Shan Chi. “Which role do non-source items play in the social sciences? A case study in political science in Germany”. In: *Scientometrics* 101.2 (Nov. 2014), pp. 1195–1213. ISSN: 1588-2861. DOI: 10.1007/s11192-014-1433-1.
- [42] Wei-Lin Chiang et al. *Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality*. Mar. 2023. URL: <https://lmsys.org/blog/2023-03-30-vicuna/>.
- [43] Peter Christen. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer Science & Business Media, 2012. DOI: 10.1007/978-3-642-31164-2.
- [44] Vassilis Christophides, Vasilis Efthymiou, and Kostas Stefanidis. “Entity Resolution in the Web of Data”. In: *Synthesis Lectures on the Semantic Web: Theory and Technology* 5.3 (2015), pp. 1–122. DOI: 10.2200/S00655ED1V01Y201507WBE013.

- [45] Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. “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)*. June 2019.
- [46] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. “SPECTER: Document-level Representation Learning using Citation-informed Transformers”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 2270–2282.
- [47] Arman Cohan and Nazli Goharian. “Scientific Article Summarization Using Citation-Context and Article’s Discourse Structure”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 390–400. DOI: 10.18653/v1/D15-1045. URL: <https://www.aclweb.org/anthology/D15-1045>.
- [48] Giovanni Colavizza and Matteo Romanello. “Citation Mining of Humanities Journals: The Progress to Date and the Challenges Ahead”. In: *Journal of European Periodical Studies* 4.1 (2019), pp. 36–53.
- [49] Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. “A Survey of Multilingual Neural Machine Translation”. In: *ACM Comput. Surv.* 53.5 (Sept. 2020). ISSN: 0360-0300. DOI: 10.1145/3406095.
- [50] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “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)*. Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.
- [51] Daniel Duma, Ewan Klein, Maria Liakata, James Ravenscroft, and Amanda Clare. “Rhetorical Classification of Anchor Text for Citation Recommendation”. In: *D-Lib Magazine* 22 (2016).
- [52] Alexander Dunn et al. *Structured information extraction from complex scientific text with fine-tuned large language models*. Dec. 2022. DOI: 10.48550/arXiv.2212.05238.
- [53] Travis Ebesu and Yi Fang. “Neural Citation Network for Context-Aware Citation Recommendation”. In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR’17. Shinjuku, Tokyo, Japan, 2017, pp. 1093–1096.
- [54] Irene Eleta and Jennifer Golbeck. “Bridging languages in social networks: How multilingual users of Twitter connect language communities?” In: *Proceedings of the American Society for Information Science and Technology* 49.1 (2012), pp. 1–4. DOI: 10.1002/meet.14504901327.

- [55] Aaron Elkiss et al. “Blind men and elephants: What do citation summaries tell us about a research article?” In: *Journal of the American Society for Information Science and Technology* 59.1 (2008), pp. 51–62. DOI: 10.1002/asi.20707.
- [56] Michael Färber. “The Microsoft Academic Knowledge Graph: A Linked Data Source with 8 Billion Triples of Scholarly Data”. In: *Proceedings of the 18th International Semantic Web Conference. ISWC’19*. Auckland, New Zealand, 2019, pp. 113–129. DOI: 10.1007/978-3-030-30796-7_8.
- [57] Michael Färber and Lin Ao. “The Microsoft Academic Knowledge Graph Enhanced: Author Name Disambiguation, Publication Classification, and Embeddings”. In: *Quantitative Science Studies* 3.1 (Apr. 2022), pp. 51–98. DOI: 10.1162/qss_a_00183.
- [58] Michael Färber and Adam Jatowt. “Citation recommendation: approaches and datasets”. In: *International Journal on Digital Libraries* 21.4 (Dec. 2020), pp. 375–405. ISSN: 1432-1300. DOI: 10.1007/s00799-020-00288-2.
- [59] Michael Färber and Ashwath Sampath. “Determining How Citations Are Used in Citation Contexts”. In: *Proceedings of the 23th International Conference on Theory and Practice of Digital Libraries. TPD L’19*. Oslo, Norway, 2019.
- [60] Michael Färber, Alexander Thiemann, and Adam Jatowt. “A High-Quality Gold Standard for Citation-based Tasks”. In: *Proceedings of the 11th International Conference on Language Resources and Evaluation. LREC’18*. Miyazaki, Japan, 2018.
- [61] Toni Feder. *Joanne Cohn and the email list that led to arXiv*. Nov. 2021. DOI: 10.1063/PT.6.4.20211108a. URL: <https://pubs.aip.org/physicstoday/online/29310/Joanne-Cohn-and-the-email-list-that-led-to-arXiv> (visited on 11/03/2023).
- [62] Ronald A Fisher. “The use of multiple measurements in taxonomic problems”. In: *Annals of eugenics* 7.2 (1936), pp. 179–188.
- [63] Yannis Foufoulas, Lefteris Stamatogiannakis, Harry Dimitropoulos, and Yannis Ioannidis. “High-Pass Text Filtering for Citation Matching”. In: *Research and Advanced Technology for Digital Libraries*. Cham: Springer International Publishing, 2017, pp. 355–366. ISBN: 978-3-319-67008-9.
- [64] Sainyam Galhotra, Donatella Firmani, Barna Saha, and Divesh Srivastava. “Efficient and effective ER with progressive blocking”. In: *The VLDB Journal* 30.4 (July 2021), pp. 537–557. ISSN: 0949-877X. DOI: 10.1007/s00778-021-00656-7.
- [65] Lukas Galke, Florian Mai, Iacopo Vagliano, and Ansgar Scherp. “Multi-Modal Adversarial Autoencoders for Recommendations of Citations and Subject Labels”. In: *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization. UMAP’18*. Singapore, Singapore: ACM, 2018, pp. 197–205. ISBN: 978-1-4503-5589-6. DOI: 10.1145/3209219.3209236.
- [66] Eugene Garfield, Irving H Sher, Richard J Torpie, et al. *The use of citation data in writing the history of science*. Tech. rep. Defense Technical Information Center, Dec. 1964.

- [67] Souvick Ghosh, Dipankar Das, and Tanmoy Chakraborty. “Determining Sentiment in Citation Text and Analyzing Its Impact on the Proposed Ranking Index”. In: *Proceedings of the 17th International Conference on Computational Linguistics and Intelligent Text Processing. CICLing’16*. Konya, Turkey, 2016, pp. 292–306.
- [68] Deyan Ginev. *arXMLiv:2020 dataset, an HTML5 conversion of arXiv.org*. hosted at <https://sigmathling.kwarc.info/resources/arxmliv-dataset-2020/>. SIGMathLing – Special Interest Group on Math Linguistics. 2020.
- [69] Paul Ginsparg. *It was twenty years ago today ...* 2011. DOI: 10.48550/arXiv.1108.2700. arXiv: 1108.2700 [cs.DL].
- [70] Bela Gipp, Norman Meuschke, and Mario Lipinski. “CITREC : An Evaluation Framework for Citation-Based Similarity Measures based on TREC Genomics and PubMed Central”. In: *iConference 2015 Proceedings*. iSchools, 2015.
- [71] Juan Gorraiz, David Melero-Fuentes, Christian Gumpenberger, and Juan-Carlos Valderrama-Zurián. “Availability of digital object identifiers (DOIs) in Web of Science and Scopus”. In: *Journal of Informetrics* 10.1 (Feb. 2016), pp. 98–109. ISSN: 1751-1577. DOI: 10.1016/j.joi.2015.11.008.
- [72] Archana Goyal, Vishal Gupta, and Manish Kumar. “Recent Named Entity Recognition and Classification techniques: A systematic review”. In: *Computer Science Review* 29 (2018), pp. 21–43. ISSN: 1574-0137. DOI: 10.1016/j.cosrev.2018.06.001.
- [73] Tudor Groza, Siegfried Handschuh, Knud Möller, and Stefan Decker. “SALT - Semantically Annotated LaTeX for Scientific Publications”. In: *The Semantic Web: Research and Applications*. Ed. by Enrico Franconi, Michael Kifer, and Wolfgang May. Springer Berlin Heidelberg, 2007, pp. 518–532. ISBN: 978-3-540-72667-8.
- [74] Venkat Gudivada, Amy Apon, and Junhua Ding. “Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations”. In: *International Journal on Advances in Software* 10 (July 2017), pp. 1–20.
- [75] Laurel L. Haak, Alice Meadows, and Josh Brown. “Using ORCID, DOI, and Other Open Identifiers in Research Evaluation”. In: *Frontiers in Research Metrics and Analytics* 3 (Oct. 2018). ISSN: 2504-0537. DOI: 10.3389/frma.2018.00028.
- [76] Scott A. Hale. “Global Connectivity and Multilinguals in the Twitter Network”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI ’14*. Toronto, Ontario, Canada: Association for Computing Machinery, 2014, pp. 833–842. ISBN: 9781450324731. DOI: 10.1145/2556288.2557203.
- [77] Scott A. Hale. “Net Increase? Cross-Lingual Linking in the Blogosphere”. In: *Journal of Computer-Mediated Communication* 17.2 (2012), pp. 135–151. DOI: 10.1111/j.1083-6101.2011.01568.x.
- [78] Corey Harper et al. “SemEval-2021 Task 8: MeasEval – Extracting Counts and Measurements and their Related Contexts”. In: *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Aug. 2021, pp. 306–316. DOI: 10.18653/v1/2021.semeval-1.38.

- [79] Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and C. Lee Giles. “Context-aware Citation Recommendation”. In: *Proceedings of the 19th International Conference on World Wide Web*. WWW’10. Raleigh, NC, USA, 2010, pp. 421–430.
- [80] Myriam Hernández-Alvarez and José M. Gomez. “Survey about citation context analysis: Tasks, techniques, and resources”. In: *Natural Language Engineering* 22.3 (2016), pp. 327–349. DOI: 10.1017/S1351324915000388.
- [81] Thomas N Herzog, Fritz J Scheuren, and William E Winkler. *Data quality and record linkage techniques*. Vol. 1. Springer, 2007.
- [82] Jorge E Hirsch. “An index to quantify an individual’s scientific research output”. In: *Proceedings of the National academy of Sciences* 102.46 (2005), pp. 16569–16572.
- [83] Wenyi Huang, Zhaohui Wu, Chen Liang, Prasenjit Mitra, and C. Lee Giles. “A Neural Probabilistic Model for Context Based Citation Recommendation”. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI’15. Austin, Texas: AAAI Press, 2015, pp. 2404–2410. ISBN: 0-262-51129-0.
- [84] Sun Huh. “Journal Article Tag Suite 1.0: National Information Standards Organization standard of journal extensible markup language”. In: *Science Editing* 1.2 (2014), pp. 99–104. DOI: 10.6087/kcse.2014.1.99.
- [85] K Hyland. “Academic attribution: citation and the construction of disciplinary knowledge”. In: *Applied Linguistics* 20.3 (1999), pp. 341–367. DOI: 10.1093/applin/20.3.341.
- [86] *ISO 32000-2:2020 (PDF 2.0) Second edition: 2020-12*. Standard. Geneva, CH: International Organization for Standardization, Dec. 2000.
- [87] Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. “SciREX: A Challenge Dataset for Document-Level Information Extraction”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, July 2020, pp. 7506–7516. DOI: 10.18653/v1/2020.acl-main.670.
- [88] Tommi Sakari Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. “Automatic language identification in texts: A survey”. In: *Journal of Artificial Intelligence Research* 65 (2019), pp. 675–782.
- [89] Zhuoren Jiang, Yao Lu, and Xiaozhong Liu. “Cross-Language Citation Recommendation via Publication Content and Citation Representation Fusion”. In: *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*. JCDL ’18. Fort Worth, Texas, USA: Association for Computing Machinery, 2018, pp. 347–348. ISBN: 9781450351782. DOI: 10.1145/3197026.3203898.
- [90] Zhuoren Jiang, Yue Yin, Liangcai Gao, Yao Lu, and Xiaozhong Liu. “Cross-Language Citation Recommendation via Hierarchical Representation Learning on Heterogeneous Graph”. In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. SIGIR ’18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 635–644. ISBN: 9781450356572. DOI: 10.1145/3209978.3210032.

- [91] Hongshan Jin, Masashi Toyoda, and Naoki Yoshinaga. “Can Cross-Lingual Information Cascades Be Predicted on Twitter?” In: *Social Informatics*. Ed. by Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha Yasseri. Cham: Springer International Publishing, 2017, pp. 457–472. ISBN: 978-3-319-67217-5.
- [92] John F. Sargent Jr. *Global Research and Development Expenditures: Fact Sheet (R44283)*. Tech. rep. R44283. Congressional Research Service, Sept. 2022. URL: <https://sgp.fas.org/crs/misc/R44283.pdf>.
- [93] Rob Johnson, Anthony Watkinson, and Michael Mabe. *The STM Report: An overview of scientific and scholarly publishing*. 2018. URL: https://www.stm-assoc.org/2018_10_04_STM_Report_2018.pdf.
- [94] J.M. Juran and A.B. Godfrey. *Juran’s Quality Handbook*. McGraw-Hill International Editions. McGraw Hill, 1999. ISBN: 9780071165396.
- [95] David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. “Measuring the Evolution of a Scientific Field through Citation Frames”. In: *Transactions of the Association for Computational Linguistics* 6 (July 2018), pp. 391–406. ISSN: 2307-387X. DOI: 10.1162/tac1_a_00028.
- [96] Charlene Kellsey and Jennifer E Knievel. “Global English in the humanities? A longitudinal citation study of foreign-language use by humanities scholars”. In: *College & Research Libraries* 65.3 (2004), pp. 194–204.
- [97] Samiya Khan, Xiufeng Liu, Kashish A. Shakil, and Mansaf Alam. “A survey on scholarly data: From big data perspective”. In: *Information Processing & Management* 53.4 (2017), pp. 923–944. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2017.03.006.
- [98] Suin Kim et al. “Understanding Editing Behaviors in Multilingual Wikipedia”. In: *PLOS ONE* 11.5 (May 2016), pp. 1–22. DOI: 10.1371/journal.pone.0155305.
- [99] Olessia Kirchik, Yves Gingras, and Vincent Larivière. “Changes in publication languages and citation practices and their effect on the scientific impact of Russian science (1993–2010)”. In: *Journal of the American Society for Information Science and Technology* 63.7 (2012), pp. 1411–1419. DOI: 10.1002/asi.22642.
- [100] Andrea Kohlhase, Michael Kohlhase, and Christoph Lange. “STEX+: a system for flexible formalization of linked data”. In: *Proceedings of the 6th International Conference on Semantic Systems*. I-SEMANTICS ’10. Association for Computing Machinery, Sept. 2010, pp. 1–9. ISBN: 9781450300148. DOI: 10.1145/1839707.1839712.
- [101] Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. “Large Language Models are Zero-Shot Reasoners”. In: *Advances in Neural Information Processing Systems* 35 (Dec. 2022), pp. 22199–22213. (Visited on 08/22/2023).
- [102] Hee-Kwan Koo et al. “Effects of Unpopular Citation Fields in Citation Matching Performance”. In: *2011 International Conference on Information Science and Applications*. 2011, pp. 1–7. DOI: 10.1109/ICISA.2011.5772372.

- [103] John E. Kratz and Carly Strasser. “Making data count”. In: *Scientific Data* 2.1 (Aug. 2015). ISSN: 2052-4463. DOI: 10.1038/sdata.2015.39.
- [104] Bernd Krieg-Brückner, Arne Lindow, Christoph Lüth, Achim Mahnke, and George Russell. “Semantic interrelation of documents via an ontology”. In: *DeLFI 2004: Die 2. e-Learning Fachtagung Informatik, Tagung der Fachgruppe e-Learning der Gesellschaft für Informatik e.V. (GI)*. Bonn: Gesellschaft für Informatik e.V., Sept. 2004, pp. 271–282. ISBN: 3-88579-381-4.
- [105] Tobias Kuhn. “A Survey and Classification of Controlled Natural Languages”. In: *Comput. Linguist.* 40.1 (Mar. 2014), pp. 121–170. ISSN: 0891-2017. DOI: 10.1162/COLI_a_00168.
- [106] Dan Lahav et al. “A Search Engine for Discovery of Scientific Challenges and Directions”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.11 (June 2022), pp. 11982–11990. DOI: 10.1609/aaai.v36i11.21456.
- [107] Viet Lai, Amir Pouran Ben Veyseh, Franck Dernoncourt, and Thien Nguyen. “SemEval 2022 Task 12: SymLink - Linking Mathematical Symbols to their Descriptions”. In: *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. July 2022. DOI: 10.18653/v1/2022.semeval-1.230.
- [108] Wout Lamers, Nees Jan van Eck, Ludo Waltman, and Holger Hoos. “Patterns in citation context: the case of the field of scientometrics”. In: *STI 2018 Conference proceedings*. Centre for Science and Technology Studies (CWTS), 2018, pp. 1114–1122.
- [109] Leslie Lamport. *LaTeX: A Document Preparation System*. Addison-Wesley, 1994. ISBN: 9780201529838.
- [110] Esther Landhuis. “Scientific literature: Information overload”. In: *Nature* 535.7612 (July 2016), pp. 457–458. ISSN: 1476-4687. DOI: 10.1038/nj7612-457a.
- [111] Daniel T. Larose and Chantal D. Larose. *Discovering Knowledge in Data: An Introduction to Data Mining*. Wiley Series on Methods and Applications in Data Mining. Wiley, 2014. ISBN: 9780470908747.
- [112] Anne Lauscher et al. “Linked Open Citation Database: Enabling Libraries to Contribute to an Open and Interconnected Citation Graph”. In: *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*. JCDL ’18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 109–118. ISBN: 9781450351782. DOI: 10.1145/3197026.3197050.
- [113] Anne Lauscher et al. *MultiCite: Modeling realistic citations requires moving beyond the single-sentence single-label setting*. 2021. arXiv: 2107.00414 [cs.CL].
- [114] Jonathan Lazar et al. “Making the field of computing more inclusive”. In: *Communications of the ACM* 60.3 (Feb. 2017), pp. 50–59. ISSN: 1557-7317. DOI: 10.1145/2993420.
- [115] Liming Liang, Ronald Rousseau, and Zhen Zhong. “Non-English journals and papers in physics and chemistry: bias in citations?” In: *Scientometrics* 95.1 (Apr. 2013), pp. 333–350. ISSN: 1588-2861. DOI: 10.1007/s11192-012-0828-0.

- [116] Andreas Liesenfeld, Alianda Lopez, and Mark Dingemans. “Opening up ChatGPT: Tracking Openness, Transparency, and Accountability in Instruction-Tuned Text Generators”. In: *Proceedings of the 5th International Conference on Conversational User Interfaces*. CUI '23. New York, NY, USA, 2023. DOI: 10.1145/3571884.3604316.
- [117] Theresa Lillis, Ann Hewings, Dimitra Vladimirou, and Mary Jane Curry. “The geolinguistics of English as an academic lingua franca: citation practices across English-medium national and English-medium international journals”. In: *International Journal of Applied Linguistics* 20.1 (2010), pp. 111–135. DOI: 10.1111/j.1473-4192.2009.00233.x.
- [118] Jialiang Lin, Yao Yu, Jiabin Song, and Xiaodong Shi. “Detecting and analyzing missing citations to published scientific entities”. In: *Scientometrics* 127.5 (May 2022), pp. 2395–2412. ISSN: 1588-2861. DOI: 10.1007/s11192-022-04334-5.
- [119] Jialiang Lin, Yao Yu, Yu Zhou, Zhiyang Zhou, and Xiaodong Shi. “How many preprints have actually been printed and why: a case study of computer science preprints on arXiv”. In: *Scientometrics* 124.1 (July 2020), pp. 555–574. ISSN: 1588-2861. DOI: 10.1007/s11192-020-03430-8.
- [120] Fang Liu, Guangyuan Hu, Li Tang, and Weishu Liu. “The penalty of containing more non-English articles”. In: *Scientometrics* 114.1 (Jan. 2018), pp. 359–366. ISSN: 1588-2861. DOI: 10.1007/s11192-017-2577-6.
- [121] Xiaomei Liu and Xiaotian Chen. “CJK Languages or English: Languages Used by Academic Journals in China, Japan, and Korea”. In: *Journal of Scholarly Publishing* 50.3 (2019), pp. 201–214.
- [122] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. “S2ORC: The Semantic Scholar Open Research Corpus”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, July 2020, pp. 4969–4983.
- [123] Patrice Lopez. “GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications”. In: *Research and Advanced Technology for Digital Libraries*. Springer Berlin Heidelberg, 2009, pp. 473–474.
- [124] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. “Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction”. In: *Proc. Conf. Empirical Methods Natural Language Process. (EMNLP)*. 2018.
- [125] Shutian Ma, Chengzhi Zhang, and Xiaozhong Liu. “A review of citation recommendation: from textual content to enriched context”. In: *Scientometrics* 122.3 (Mar. 2020), pp. 1445–1472. ISSN: 1588-2861.
- [126] Yuning Mao, Ming Zhong, and Jiawei Han. “CiteSum: Citation Text-guided Scientific Extreme Summarization and Domain Adaptation with Limited Supervision”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Dec. 2022, pp. 10922–10935. URL: <https://aclanthology.org/2022.emnlp-main.750>.

- [127] Alberto Martín-Martín, Mike Thelwall, Enrique Orduna-Malea, and Emilio Delgado López-Cózar. “Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations’ COCI: a multidisciplinary comparison of coverage via citations”. In: *Scientometrics* 126.1 (2021), pp. 871–906. ISSN: 1588-2861. DOI: 10.1007/s11192-020-03690-4.
- [128] Alice Meadows, Laurel L. Haak, and Josh Brown. “Persistent identifiers: the building blocks of the research information infrastructure”. In: *Insights* 32 (2019), pp. 1–6. DOI: 10.1629/uksg.457.
- [129] Karine Megerdooian and Dan Parvaz. “Low-Density Language Bootstrapping: the Case of Tajiki Persian”. In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*. May 2008. URL: http://www.lrec-conf.org/proceedings/lrec2008/pdf/827_paper.pdf.
- [130] Xiangyi Meng, Onur Varol, and Albert-László Barabási. “Hidden Citations Obscure True Impact in Science”. In: *Proceedings of the International Conference on Science of Science and Innovation 2023*. June 2023.
- [131] Dominique Mercier, Syed Rizvi, Vikas Rajashekar, Andreas Dengel, and Sheraz Ahmed. “ImpactCite: An XLNet-based Solution Enabling Qualitative Citation Impact Analysis Utilizing Sentiment and Intent”. In: *Proceedings of the 13th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART, INSTICC*. SciTePress, 2021, pp. 159–168. ISBN: 978-989-758-484-8. DOI: 10.5220/0010235201590168.
- [132] Melina Meyer, Jenny Frey, Tamino Laub, Marco Wrzalik, and Dirk Krechel. “Citcom – Citation Recommendation”. In: *INFORMATIK 2020*. Ed. by Ralf H. Reussner, Anne Koziol, and Robert Heinrich. Gesellschaft für Informatik, Bonn, 2021, pp. 907–914. DOI: 10.18420/inf2020_82.
- [133] John Mingers and Loet Leydesdorff. “A review of theory and practice in scientometrics”. In: *European Journal of Operational Research* 246.1 (2015), pp. 1–19. ISSN: 0377-2217. DOI: 10.1016/j.ejor.2015.04.002.
- [134] Frank Mittelbach and Ulrike Fischer. *The LaTeX Companion: Parts I & II, 3rd Edition*. Addison Wesley, 2023. ISBN: 9780138166489.
- [135] Frank Mittelbach and Chris Rowley. “LATEX Tagged PDF—A blueprint for a large project”. In: *TUG-boat* 41.3 (2020), pp. 292–298.
- [136] Daichi Mochihashi. “Researcher2Vec: Neural Linear Model of Scholar Recommendation for Funding Agency”. In: *Proceedings of the 19th International Conference of the International Society for Scientometrics and Informetrics*. July 2023. DOI: 10.5281/zenodo.8428813.
- [137] Henk F. Moed, Valentina Markusova, and Mark Akoev. “Trends in Russian research output indexed in Scopus and Web of Science”. In: *Scientometrics* 116.2 (Aug. 2018), pp. 1153–1180. ISSN: 1588-2861.

- [138] Saif Mohammad et al. “Using Citations to Generate surveys of Scientific Paradigms”. In: *Proceedings of the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. NAACL-HLT’09. Boulder, Colorado, USA, 2009, pp. 584–592.
- [139] Dattatreya Mohapatra, Abhishek Maiti, Sumit Bhatia, and Tanmoy Chakraborty. “Go Wide, Go Deep: Quantifying the Impact of Scientific Papers Through Influence Dispersion Trees”. In: *Proceedings of the 19th ACM/IEEE Joint Conference on Digital Libraries*. JCDL’19. Champaign, IL, USA, 2019, pp. 305–314.
- [140] Scott L. Montgomery. *Does science need a global language?: English and the future of research*. University of Chicago Press, 2013. ISBN: 9780226535036.
- [141] Michael J Moravcsik and Poovanalingam Murugesan. “Some results on the function and quality of citations”. In: *Social studies of science* 5.1 (1975), pp. 86–92.
- [142] Olga Moskaleva and Mark Akoev. “Non-English language publications in Citation Indexes - quantity and quality”. In: *Proceedings 17th International Conference on Scientometrics & Informetrics*. Vol. 1. Italy: Edizioni Efesto, Sept. 2019, pp. 35–46. ISBN: 978-88-3381-118-5.
- [143] 英嗣 難波. 論文間の参照情報を考慮したサーベイ論文作成支援システムの開発. Feb. 1998. URL: <http://hdl.handle.net/10119/1151>. [Hidetsugu Nanba, Multi-Paper Summarization Using Reference Information (in Japanese)].
- [144] Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. “Information extraction from scientific articles: a survey”. In: *Scientometrics* 117.3 (Dec. 2018), pp. 1931–1990. ISSN: 1588-2861. DOI: 10.1007/s11192-018-2921-5.
- [145] Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. “Named Entity Recognition and Relation Extraction: State-of-the-Art”. In: *ACM Comput. Surv.* 54.1 (Feb. 2021). ISSN: 0360-0300. DOI: 10.1145/3445965.
- [146] OECD. *Main Science and Technology Indicators, Volume 2022 Issue 2*. Tech. rep. June 2023. DOI: 10.1787/1cdcb031-en.
- [147] George Papadakis, Ekaterini Ioannou, Claudia Niederee, and Peter Fankhauser. “Efficient Entity Resolution for Large Heterogeneous Information Spaces”. In: May 2011, pp. 535–544. DOI: 10.1145/1935826.1935903.
- [148] George Papadakis, Georgia Koutrika, Themis Palpanas, and Wolfgang Nejdl. “Meta-Blocking: Taking Entity Resolution to the Next Level”. In: *IEEE Transactions on Knowledge and Data Engineering* 26 (Aug. 2014). DOI: 10.1109/TKDE.2013.54.
- [149] George Papadakis, Dimitrios Skoutas, Emmanouil Thanos, and Themis Palpanas. “Blocking and Filtering Techniques for Entity Resolution: A Survey”. In: *ACM Computing Surveys* 53.2 (Mar. 2020), 31:1–31:42. ISSN: 0360-0300. DOI: 10.1145/3377455.
- [150] George Papadakis, Jonathan Svirsky, Avigdor Gal, and Themis Palpanas. “Comparative Analysis of Approximate Blocking Techniques for Entity Resolution”. In: *Proc. VLDB Endow.* 9.9 (May 2016), pp. 684–695. ISSN: 2150-8097. DOI: 10.14778/2947618.2947624.

- [151] Mathias Parisot and Jakub Zavrel. “Multi-objective Representation Learning for Scientific Document Retrieval”. In: *Proceedings of the Third Workshop on Scholarly Document Processing*. Gyeongju, Republic of Korea: Association for Computational Linguistics, Oct. 2022, pp. 80–88. URL: <https://aclanthology.org/2022.sdp-1.9>.
- [152] Krutarth Patel, Cornelia Caragea, Doina Caragea, and C. Lee Giles. “Author Homepage Discovery in CiteSeerX”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.17 (May 2021), pp. 15146–15155. DOI: 10.1609/aaai.v35i17.17778.
- [153] Maciej P. Polak and Dane Morgan. *Extracting Accurate Materials Data from Research Papers with Conversational Language Models and Prompt Engineering – Example of ChatGPT*. Mar. 2023. DOI: 10.48550/arXiv.2303.05352.
- [154] Nancy Pontika, Petr Knoth, Matteo Cancellieri, and Samuel Pearce. “Developing Infrastructure to Support Closer Collaboration of Aggregators with Open Repositories”. In: *LIBER Quarterly* 25.4 (Apr. 2016), pp. 172–188. URL: <http://oro.open.ac.uk/45935/>.
- [155] Animesh Prasad, Manpreet Kaur, and Min-Yen Kan. “Neural ParsCit: A Deep Learning Based Reference String Parser”. In: *International Journal on Digital Libraries* 19 (2018), pp. 323–337. DOI: 10.1007/s00799-018-0242-1.
- [156] Jason Priem, Heather Piwowar, and Richard Orr. *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*. 2022. DOI: 10.48550/ARXIV.2205.01833. URL: <https://arxiv.org/abs/2205.01833>.
- [157] Behrang QasemiZadeh and Anne-Kathrin Schumann. “The ACL RD-TEC 2.0: A Language Resource for Evaluating Term Extraction and Entity Recognition Methods”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*. European Language Resources Association (ELRA), May 2016, pp. 1862–1868.
- [158] Daniela Raciti, Karen Yook, Todd W Harris, Tim Schedl, and Paul W Sternberg. “Micropublication: incentivizing community curation and placing unpublished data into the public domain”. In: *Database* 2018 (Jan. 2018). ISSN: 1758-0463. DOI: 10.1093/database/bay013.
- [159] Dragomir R. Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. “The ACL anthology network corpus”. In: *Language Resources and Evaluation* 47.4 (2013), pp. 919–944.
- [160] Edward Raff. “A Step Toward Quantifying Independently Reproducible Machine Learning Research”. In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019.
- [161] Ingkarat Rak-Amnouykit, Ana Milanova, Guillaume Baudart, Martin Hirzel, and Julian Dolby. “Extracting Hyperparameter Constraints from Code”. In: *ICLR Workshop on Security and Safety in Machine Learning Systems*. May 2021. URL: <https://hal.science/hal-03401683> (visited on 12/16/2023).

- [162] Krithika Ramesh, Sunayana Sitaram, and Monojit Choudhury. “Fairness in Language Models Beyond English: Gaps and Challenges”. In: *Findings of the Association for Computational Linguistics: EACL 2023*. Ed. by Andreas Vlachos and Isabelle Augenstein. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 2106–2119. DOI: 10.18653/v1/2023.findings-eacl.157.
- [163] Yaniv Reingewertz and Carmela Lutmar. “Academic in-group bias: An empirical examination of the link between author and journal affiliation”. In: *Journal of Informetrics* 12.1 (2018), pp. 74–86. ISSN: 1751-1577. DOI: 10.1016/j.joi.2017.11.006.
- [164] Dwaipayan Roy, Kunal Ray, and Mandar Mitra. “From a Scholarly Big Dataset to a Test Collection for Bibliographic Citation Recommendation”. In: *AAAI Workshops*. 2016. URL: <https://www.aaai.org/ocs/index.php/WS/AAAIW16/paper/view/12635>.
- [165] Anna Samoilenko, Fariba Karimi, Daniel Edler, Jérôme Kunegis, and Markus Strohmaier. “Linguistic neighbourhoods: explaining cultural borders on Wikipedia through multilingual co-editing activity”. In: *EPJ Data Science* 5.1 (Mar. 2016), p. 9. ISSN: 2193-1127.
- [166] Thomas Scheidsteger and Robin Haunschild. *Comparison of metadata with relevance for bibliometrics between Microsoft Academic Graph and OpenAlex until 2020*. 2022. DOI: 10.48550/arXiv.2206.14168. URL: <https://arxiv.org/abs/2206.14168>.
- [167] Bess Schrader. *Cross-language Citation Analysis of Traditional and Open Access Journals*. Feb. 2019. DOI: 10.17615/djpr-1k06.
- [168] Athar Sefid. “Record Linkage Between CiteSeerX and Scholarly Big Datasets”. MA thesis. The Pennsylvania State University, 2019.
- [169] Akshay Sethi, Anush Sankaran, Naveen Panwar, Shreya Khare, and Senthil Mani. “DLPaper2Code: Auto-Generation of Code From Deep Learning Research Papers”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1 (Apr. 2018). DOI: 10.1609/aaai.v32i1.12326.
- [170] Fei Shu, Charles-Antoine Julien, and Vincent Larivière. “Does the web of science accurately represent chinese scientific performance?” In: *Journal of the Association for Information Science and Technology* 70.10 (2019), pp. 1138–1152. DOI: doi.org/10.1002/asi.24184.
- [171] Giovanni Simonini, Sonia Bergamaschi, and H. V. Jagadish. “BLAST: A Loosely Schema-Aware Meta-Blocking Approach for Entity Resolution”. In: *Proc. VLDB Endow.* 9.12 (Aug. 2016), pp. 1173–1184. ISSN: 2150-8097. DOI: 10.14778/2994509.2994533.
- [172] Giovanni Simonini, George Papadakis, Themis Palpanas, and Sonia Bergamaschi. “Schema-Agnostic Progressive Entity Resolution”. In: *IEEE Transactions on Knowledge and Data Engineering* 31.6 (June 2019), pp. 1208–1221. ISSN: 1558-2191. DOI: 10.1109/TKDE.2018.2852763.

- [173] Arnab Sinha et al. “An Overview of Microsoft Academic Service (MAS) and Applications”. In: *Proceedings of the 24th International Conference on World Wide Web. WWW '15 Companion*. ACM, 2015, pp. 243–246. ISBN: 978-1-4503-3473-0. DOI: 10.1145/2740908.2742839.
- [174] Heinrich Stamerjohanns and Michael Kohlhase. “Transforming the arXiv to XML”. In: *Intelligent Computer Mathematics. Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer, 2008, pp. 574–582. ISBN: 9783540851103. DOI: 10.1007/978-3-540-85110-3_46.
- [175] Markus Stocker et al. “FAIR scientific information with the Open Research Knowledge Graph”. In: *FAIR Connect 1.1 (2023)*, pp. 19–21. ISSN: 2949-799X. DOI: 10.3233/FC-221513.
- [176] Diane M. Strong, Yang W. Lee, and Richard Y. Wang. “Data Quality in Context”. In: *Commun. ACM* 40.5 (May 1997), pp. 103–110. ISSN: 0001-0782. DOI: 10.1145/253769.253804.
- [177] Andreas Strotmann and Dangzhi Zhao. “An 80/20 Data Quality Law for Professional Scientometrics?” In: *Proceedings of ISSI 2015 Istanbul: 15th International Society of Scientometrics and Informetrics Conference*. June 2015. URL: <http://www.issi2015.org/files/downloads/all-papers/1218.pdf>.
- [178] Jianlin Su et al. “Roformer: Enhanced transformer with rotary position embedding”. In: *arXiv preprint arXiv:2104.09864* (2021).
- [179] Kazunari Sugiyama and Min-Yen Kan. “A Comprehensive Evaluation of Scholarly Paper Recommendation Using Potential Citation Papers”. In: *International Journal on Digital Libraries* 16.2 (2015), pp. 91–109.
- [180] John Swales. *Genre analysis: English in academic and research settings*. Cambridge University Press, 1990.
- [181] Xuewei Tang, Xiaojun Wan, and Xun Zhang. “Cross-Language Context-Aware Citation Recommendation in Scientific Articles”. In: *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. SIGIR '14*. Association for Computing Machinery, 2014, pp. 817–826. ISBN: 9781450322577. DOI: 10.1145/2600428.2609564.
- [182] Ross Taylor et al. *GALACTICA: A Large Language Model for Science*. 2022.
- [183] Simone Teufel, Advaith Siddharthan, and Dan Tidhar. “An Annotation Scheme for Citation Function”. In: *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue. SigDIAL '06*. Association for Computational Linguistics, 2006, pp. 80–87. ISBN: 1-932432-71-X.
- [184] Simone Teufel, Advaith Siddharthan, and Dan Tidhar. “Automatic classification of citation function”. In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. EMNLP'06*. Sydney, Australia, 2006, pp. 103–110.

- [185] Dominika Tkaczyk, Andrew Collins, Paraic Sheridan, and Joeran Beel. “Machine Learning vs. Rules and Out-of-the-Box vs. Retrained: An Evaluation of Open-Source Bibliographic Reference and Citation Parsers”. In: *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*. JCDL ’18. Fort Worth, Texas, USA: ACM, 2018, pp. 99–108. ISBN: 978-1-4503-5178-2. DOI: 10.1145/3197026.3197048.
- [186] Dominika Tkaczyk, Pawel Szostek, Mateusz Fedoryszak, Pitor Jan Dendek, and Lukasz Bolikowski. “CERMINE: automatic extraction of structured metadata from scientific literature”. In: *International Journal on Document Analysis and Recognition (IJ DAR)* 18.4 (Dec. 2015), pp. 317–335. ISSN: 1433-2825.
- [187] Marco Valenzuela, Vu Ha, and Oren Etzioni. “Identifying Meaningful Citations”. In: *AAAI Workshops*. 2015. URL: <https://www.aaai.org/ocs/index.php/WS/AAAIW15/paper/view/10185>.
- [188] Michele Delli Veneri et al. “How Have Astronomers Cited Other Fields in the Last Decade?” In: *Research Notes of the AAS* 6.6 (June 2022), p. 113. DOI: 10.3847/2515-5172/ac74c7.
- [189] Miguel-Angel Vera-Baceta, Michael Thelwall, and Kayvan Kousha. “Web of Science and Scopus language coverage”. In: *Scientometrics* 121.3 (Dec. 2019), pp. 1803–1813. ISSN: 1588-2861.
- [190] Vijay Viswanathan, Graham Neubig, and Pengfei Liu. “CitationIE: Leveraging the Citation Graph for Scientific Information Extraction”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Aug. 2021, pp. 719–731. DOI: 10.18653/v1/2021.acl-long.59.
- [191] David Wadden et al. “Fact or Fiction: Verifying Scientific Claims”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Nov. 2020, pp. 7534–7550. DOI: 10.18653/v1/2020.emnlp-main.609.
- [192] Kuansan Wang et al. “A Review of Microsoft Academic Services for Science of Science Studies”. In: *Frontiers in Big Data* 2 (2019), p. 45. ISSN: 2624-909X. DOI: 10.3389/fdata.2019.00045.
- [193] Shuhe Wang et al. *GPT-NER: Named Entity Recognition via Large Language Models*. May 2023. DOI: 10.48550/arXiv.2304.10428.
- [194] Michael Whidby, David Zajic, and Bonnie Dorr. *Citation handling for improved summarization of scientific documents*. Tech. rep. 2011.
- [195] Mark D. Wilkinson et al. “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific Data* 3.1 (2016), p. 160018. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18.
- [196] Jian Wu, Ryan Hiltabrand, Dominik Soós, and C. Lee Giles. “Scholarly Big Data Quality Assessment: A Case Study of Document Linking and Conflation with S2ORC”. In: *Proceedings of the 22nd ACM Symposium on Document Engineering*. DocEng ’22. New York, NY, USA: Association for Computing Machinery, 2022. ISBN: 9781450395441. DOI: 10.1145/3558100.3563850.

- [197] Jian Wu, Kunho Kim, and C. Lee Giles. “CiteSeerX: 20 Years of Service to Scholarly Big Data”. In: *Proceedings of the Conference on Artificial Intelligence for Data Discovery and Reuse. AIDR '19*. 2019. ISBN: 9781450371841. DOI: 10.1145/3359115.3359119.
- [198] Jian Wu, Chen Liang, Huaiyu Yang, and C. Lee Giles. “CiteSeerX Data: Semanticizing Scholarly Papers”. In: *Proceedings of the International Workshop on Semantic Big Data. SBD '16*. New York, NY, USA: Association for Computing Machinery, 2016. ISBN: 9781450342995. DOI: 10.1145/2928294.2928306.
- [199] Jian Wu et al. “CiteSeerX: AI in a Digital Library Search Engine”. In: *AI Magazine* 36.3 (Sept. 2015), pp. 35–48. DOI: 10.1609/aimag.v36i3.2601.
- [200] Tong Xie et al. *Large Language Models as Master Key: Unlocking the Secrets of Materials Science with GPT*. Apr. 2023. DOI: 10.48550/arXiv.2304.02213.
- [201] Can Xu et al. *WizardLM: Empowering Large Language Models to Follow Complex Instructions*. 2023.
- [202] Jiayue Xue. “An analysis of the semantic shifts of citations”. MA thesis. University of Helsinki, Faculty of Science, June 2021. URL: <http://urn.fi/URN:NBN:fi:hulib-202107263435>.
- [203] Jingfeng Yang et al. *Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond*. Apr. 2023. DOI: 10.48550/arXiv.2304.13712.
- [204] Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. “Packed Levitated Marker for Entity and Relation Extraction”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, May 2022, pp. 4904–4917. DOI: 10.18653/v1/2022.acl-long.337.
- [205] Jan Youtie, Stephen Carley, Alan L. Porter, and Philip Shapira. “Tracking researchers and their outputs: new insights from ORCID”. In: *Scientometrics* 113.1 (July 2017), pp. 437–453. ISSN: 1588-2861. DOI: 10.1007/s11192-017-2473-0.
- [206] Xinyan Yu, Trina Chatterjee, Akari Asai, Junjie Hu, and Eunsol Choi. “Beyond Counting Datasets: A Survey of Multilingual Dataset Construction and Necessary Resources”. In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 3725–3743. DOI: 10.18653/v1/2022.findings-emnlp.273.
- [207] Ethan Zuckerman. “Meet the bridgebloggers”. In: *Public Choice* 134.1 (Jan. 2008), pp. 47–65. ISSN: 1573-7101.

A

Appendix

A.1. Geographic Origin of All Cited Non-English Languages

In Figure A.1 we show the geographic origin of cross-lingual citations in relative terms per cited language (i.e., the numbers of each *row* add up to 1). The distinct diagonal of the matrix and the horizontal line for affiliations in English-speaking countries reflect the fact that most cross-lingual citations are either to a local language or originate from an English-speaking country. Among cited languages with a low number of total occurrences we can furthermore see a few cases showing unusual distributions, such as a single citation to Macedonian from an author affiliated with a Polish institution, or citations to Icelandic, where a single one originates from Iceland, while the remaining nine originate from institutions in countries where Japanese (3), Italian (1), and Swedish (5) are the most common language.

A.2. Citation Intent and Sentiment Classification

For the model training of both citation intent classification and citation sentiment classification, we fine-tune SciBERT uncased¹ using the following model configuration shown in Table A.1.

For determining the citation intent, we use the train, validation, and test split provided by the SciCite data set² (train: 74%, val: 8.3%, test: 16.9%). For citation sentiment, we split the Athar data set into train, validation, and test sets into 80%, 10%, and 10%, respectively.

¹ See https://huggingface.co/allenai/scibert_scivocab_uncased [last accessed: 2024-01-10].

² See <https://huggingface.co/datasets/scicite> [last accessed: 2024-01-10].

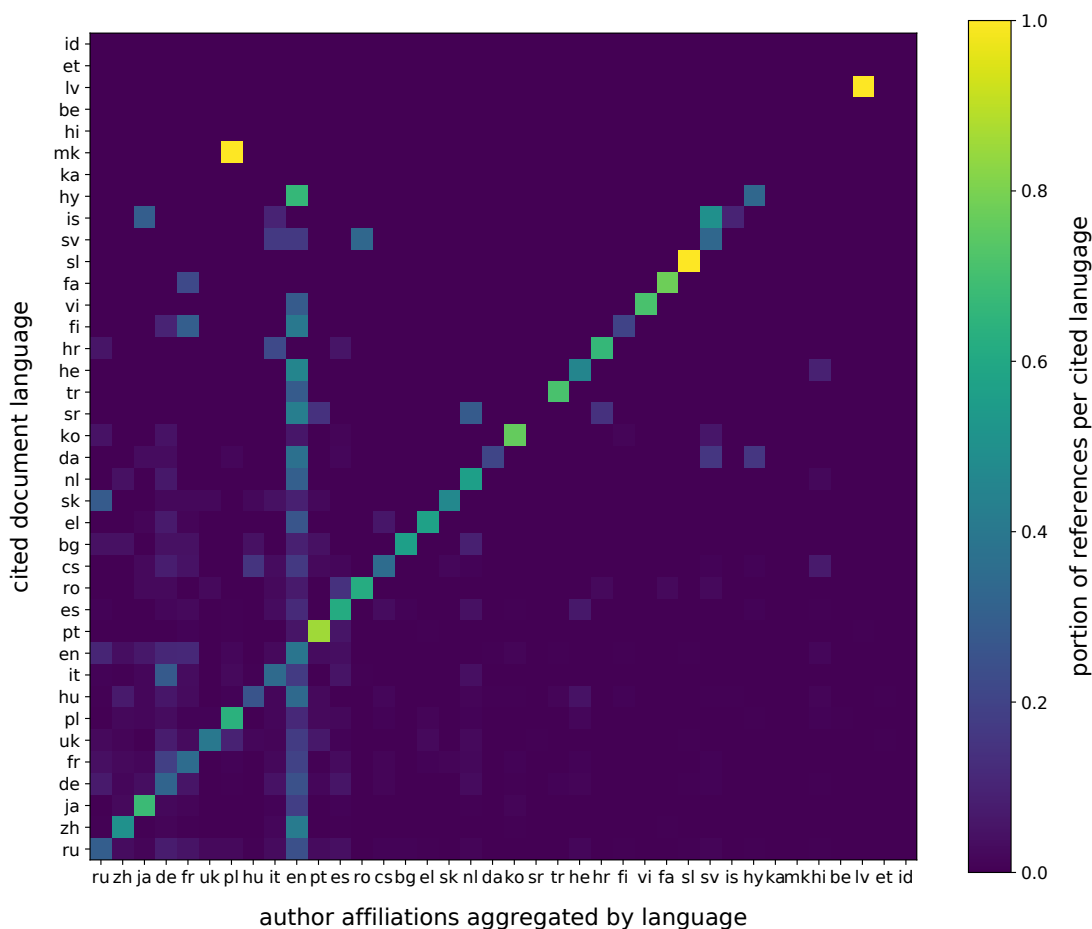


Figure A.1.: Geographic origin of cross-lingual citations (relative count).

A.3. HyperPIE Implementation Details

Fine-Tuned Models: We obtain the source code of PL-Marker from the author’s GitHub repository³. To make it work with our entity and relation schema, we extended the source code in `run_acener.py` and `run_re.py`. A patch file with all changes is provided in our code share. Our own RE model is a FFNN implemented with 4 hidden layers, each with ReLU activation and dimensions 300, 100, 25 and 2 respectively. All fine-tuned models are trained and evaluated on a local server with a GeForce RTX 3090 (24 GB).

LLMs: GPT-3.5 was accessed through the official API. The total usage cost for all testing, prompt tuning, and the full evaluation runs sums up to 60 USD. In zero-shot setting, all open models are run on a high performance compute cluster using the API layer Basaran.⁴

³ See <https://github.com/thunlp/PL-Marker/> [last accessed: 2024-01-10].

⁴ See <https://github.com/hyperonym/basaran/> [last accessed: 2024-01-10].

Table A.1.: Model configuration used for training.

Hyperparameter	value
attention_probs_dropout_prob	0.1
gradient_checkpointing	false
hidden_act	gelu
hidden_dropout_prob	0.1
hidden_size	768
initializer_range	0.02
intermediate_size	3072
layer_norm_eps	1e-12
max_position_embeddings	512
model_type	bert
num_attention_heads	12
num_hidden_layers	12
pad_token_id	0
position_embedding_type	absolute
transformers_version	4.4.2
type_vocab_size	2
use_cache	true
vocab_size	31090

Vicuna and WizardLM are run on nodes with 4× NVIDIA Tesla V100 (32 GB). GALACTICA and Falcon are run with half precision on nodes with 4× NVIDIA A100 (80 GB).

A zero-shot prompt example is shown in Listing A.1.

```

In the context of machine learning and related fields, what (if any) are the entities (datasets, models,
methods, loss functions, regularization techniques) mentioned in the LaTeX Input Text below? What (if any) are
their parameters and values?

[LaTeX Input Text start]
We use AdamW with a learning rate ( $\alpha$ ) of 1e-3 for /* [...] */
[LaTeX Input Text end]

Answer in the following YAML format.

Format:
---
text_contains_entities: true/false
entities:
  - entity<N>:
      id: e<N>
      name: "<entity name>"
      type: dataset/model/method/loss function/regularization technique
      has_parameters: true/false
      parameters:
        - parameter<M>:
            id: p<N.M>
/* [...] */
...

Only include entities that are of type dataset, model, method, loss function, or regularization technique. Do
not output entities that are of another type. Do not include entities of type task, metric, library, software,
or API.
Only produce output in the YAML format specified above. Output no additional text.

Output:

```

Listing A.1: Zero-shot prompt example.

For few-shot prompting, we employed 4 bit quantization and used the llama-cpp-python⁵ API. We use the default generation setup in llama.cpp with parameters: temperature = 0, half precision = enabled, and repetition penalty = 1.1. A few-shot prompt example is shown in Listing A.2.

```

### Instruction:
In the context of machine learning and related fields, what (if any) are the entities (datasets, models,
methods, loss functions, regularization techniques) mentioned in the LaTeX Input Text below? What (if any) are
their parameters and values?

Answer in the following YAML format.

Format:
'''
has_entities: true/false
entities:
  - entity<N>:
      id: e<N>
      name: "<entity name>"
      type: dataset/model/method/loss function/regularization technique
      has_parameters: true/false
      parameters:
        - parameter<M>:
            id: p<N.M>
/* [...] */
'''

Here are several examples.

### Example 1:

[LaTeX Input Text start]
We use AdamW with a learning rate ( $\alpha$ ) of 1e-3 for /* [...] */
[LaTeX Input Text end]

### Response 1:
'''
has_entities: true
  - entity1:
      id: e1
      name: "AdamW"
      has_parameters: true
      parameters:
        - parameter1:
            id: p1
/* [...] */
'''

### Example 2:

[LaTeX Input Text start]
/* [...] */
[LaTeX Input Text end]

### Response 2:
'''
/* [...] */
'''

### Example 3:

[LaTeX Input Text start]
/* [...] */
[LaTeX Input Text end]

### Response 3:
'''
/* [...] */
'''

Only include entities that are of type dataset, model, method, loss function, or regularization technique. Do
not output entities that are of another type. Do not include entities of type task, metric, library, software,
or API.
Only produce output in the YAML format specified above. Output no additional text.

[LaTeX Input Text start]
We use AdamW with a learning rate ( $\alpha$ ) of 1e-3 for /* [...] */
[LaTeX Input Text end]

### Response:
'''

```

Listing A.2: Few-shot prompt example.

⁵ See <https://github.com/abetlen/llama-cpp-python/> [last accessed: 2024-01-10].

In the examples given in the few-shot prompts, we omitted the field type: `dataset/model/method/loss function/regularization technique`, because this information is not part of the gold annotation. As a consequence, the model outputs also tend to skip this attribute.