

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

Dissertation Defense

Tarek Saier | 22. April 2024



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

**Data Mining and Information Extraction Methods
for Large-Scale High Quality
Scholarly Data**

Scholarly Data

■ Usage

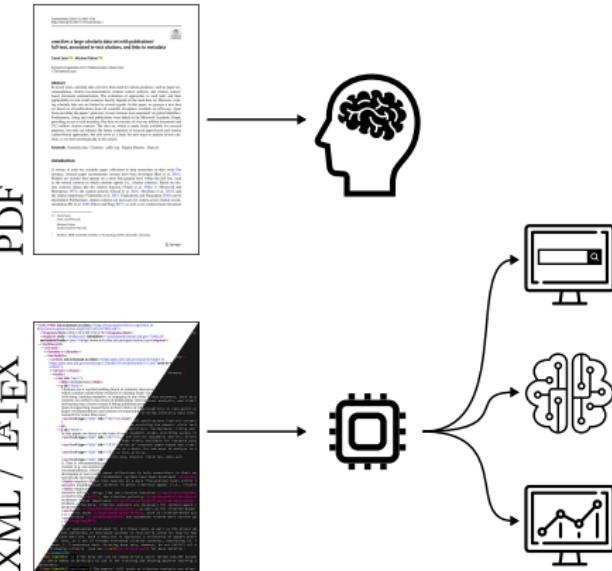
- **Services** (search, recommendation, statistics)
 - **ML models** (LLMs, summarization, recommender systems)
 - **Analyses** (temporal, geographic, institutional)

■ Flavors

- **Metadata** (MAG, OpenAlex, ORKG, crossref)
 - **Documents** (Core, arXMLiv, PMC)
 - **Linked Documents** (unarXive, S2ORC)

■ Data Sources

- PDF (Core, ACL Anthology)
 - XML (PubMed, PLOS, publisher internal)
 - LaTeX (arXiv)



Motivation



Background

Outline

Corpus

Citations & Non-English ooo

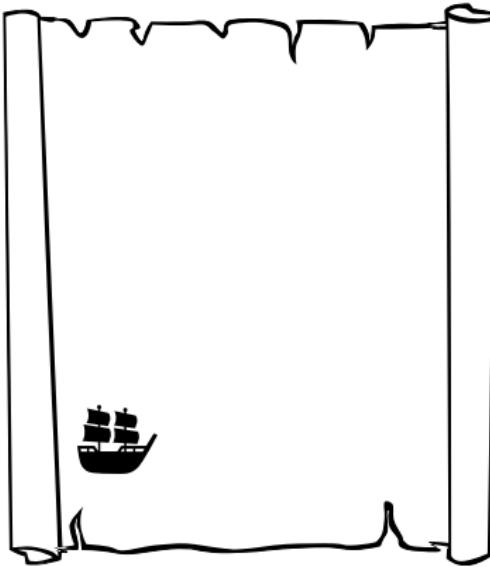
Artifact Parameters

Conclusion

References

Analogy

Maps of the Sea



Motivation



Background

Outline

- oo

Corpus



Citations & Non-English ooo

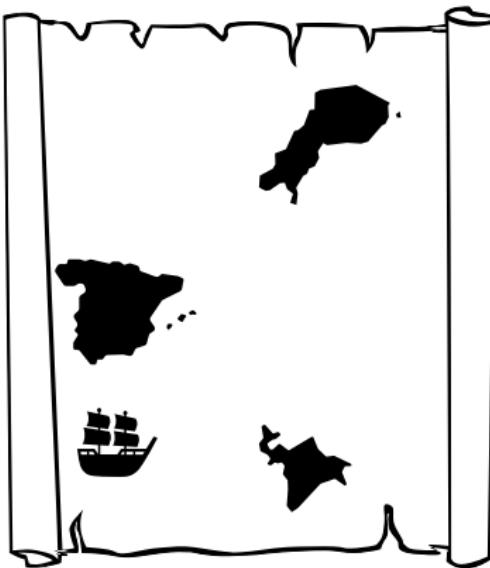
Artifact Parameters

oooooooooooooooo

Conclusion

References

Maps of the Sea



Motivation



Background

Outline

Corpus

oooooooooooo

Citations & Non-English ooo

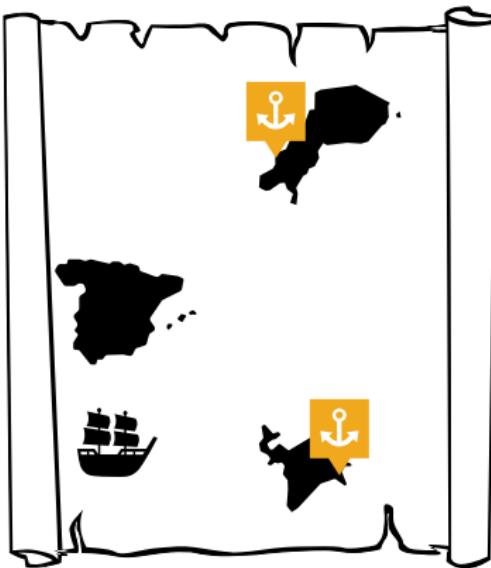
Artifact Parameters

oooooooooooooooo

Conclusion

References

Maps of the Sea



Motivation



Background

Outline

Corpus



Citations & Non-English ooo

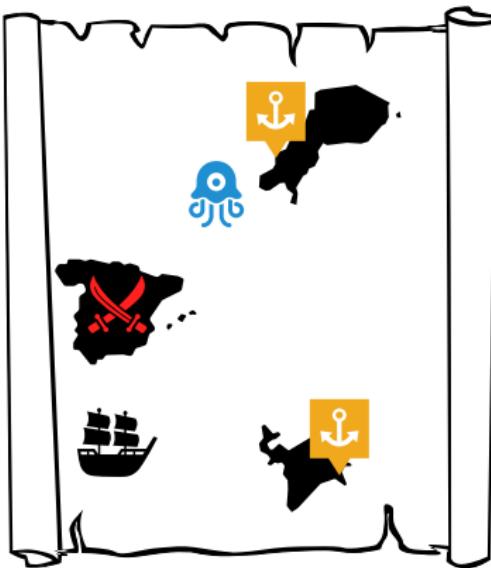
Artifact Parameters

oooooooooooooooo

Conclusion

References

Maps of the Sea



Motivation



Background

Outline

Corpus



Citations & Non-English ooo

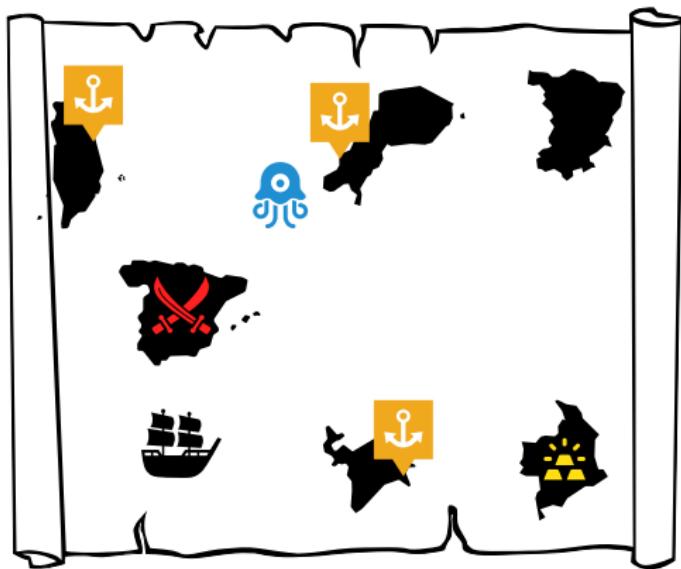
Artifact Parameters

oooooooooooooooo

Conclusion

References

Maps of the Sea



Motivation



Background

Outline

Corpus



Citations & Non-English ooo

Artifact Parameters

oooooooooooooooo

Conclusion

References

Maps of the Sea / Maps of Science

The Sailor looks for

- Port to trade
- Island to explore

The Scientist looks for

- Paper to read
- Venue to publish at
- Research idea to explore

Motivation
oooo●

Background
oooo

Outline
oo

Corpus
oooooooo

Citations & Non-English
ooo

Artifact Parameters
oooooooooooo

Conclusion
oooooo

References

Maps of the Sea / Maps of Science

The Trade Company looks for

- Routes to expand
- Ports to build
- Sailor to hire

The University/Funding Body looks for

- Research to fund
- Researcher to hire
- Policy to establish

Motivation
oooo●

Background
oooo

Outline
oo

Corpus
oooooooo

Citations & Non-English
ooo

Artifact Parameters
oooooooooooo

Conclusion
ooooo

References

Maps of the Sea / Maps of Science

The Trade Company looks for

- Routes to expand
- Ports to build
- Sailor to hire

The University/Funding Body looks for

- Research to fund
- Researcher to hire
- Policy to establish

better maps ⇒ better decisions

Motivation
oooo●

Background
ooo○

Outline
oo

Corpus
oooooooo

Citations & Non-English
ooo

Artifact Parameters
oooooooooooo

Conclusion
oooo○○

References

Maps of the Sea / Maps of Science

The Trade Company looks for

- Routes to expand
- Ports to build
- Sailor to hire

The University/Funding Body looks for

- Research to fund
- Researcher to hire
- Policy to establish

false maps ⇒ misleading/false analyses, models, etc.

Maps of the Sea / Maps of Science

The Trade Company looks for

- Routes to expand
- Ports to build
- Sailor to hire

The University/Funding Body looks for

- Research to fund
- Researcher to hire
- Policy to establish

our maps of science are insufficient

Motivation
oooo●

Background
oooo

Outline
oo

Corpus
oooooooo

Citations & Non-English
ooo

Artifact Parameters
oooooooooooo

Conclusion
oooooo

References

Maps of the Sea / Maps of Science

The Trade Company looks for

- Routes to expand
- Ports to build
- Sailor to hire

The University/Funding Body looks for

- Research to fund
- Researcher to hire
- Policy to establish



Research Objective

Develop methods for generating large-scale, high quality scholarly data.

Motivation
oooo●

Background
oooo

Outline
oo

Corpus
oooooooo

Citations & Non-English
ooo

Artifact Parameters
oooooooooooo

Conclusion
oooooo

References

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

Maps of Science



Motivation
ooooo

Background
○●○○

Outline
○○

Corpus
oooooooo

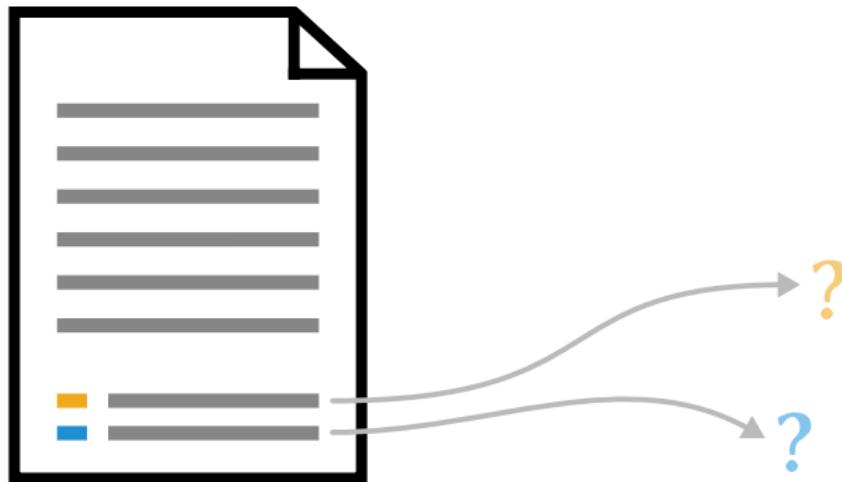
Citations & Non-English
ooo

Artifact Parameters
oooooooooooo

Conclusion
oooooo

References

Maps of Science



Motivation
ooooo

Background
○●○○

Outline
○○

Corpus
oooooooo

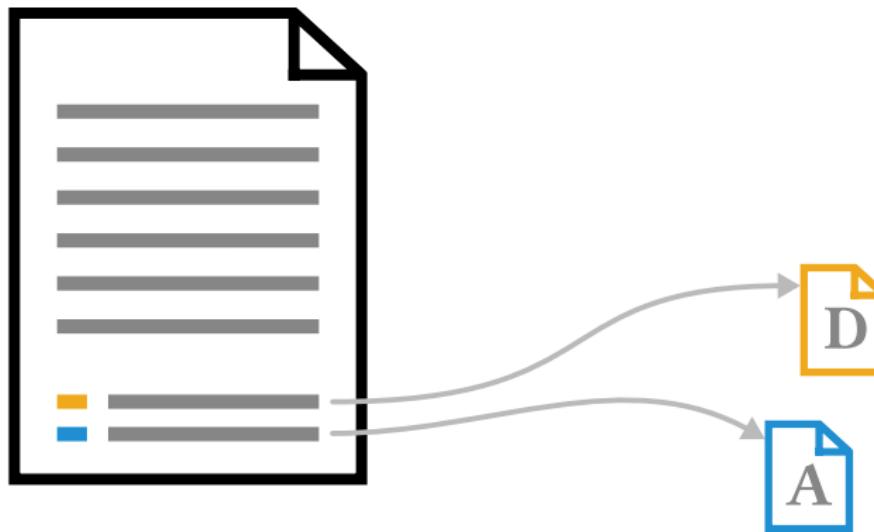
Citations & Non-English
ooo

Artifact Parameters
oooooooooooo

Conclusion
oooooo

References

Maps of Science



Motivation
ooooo

Background
○●○○

Outline
○○

Corpus
oooooooooo

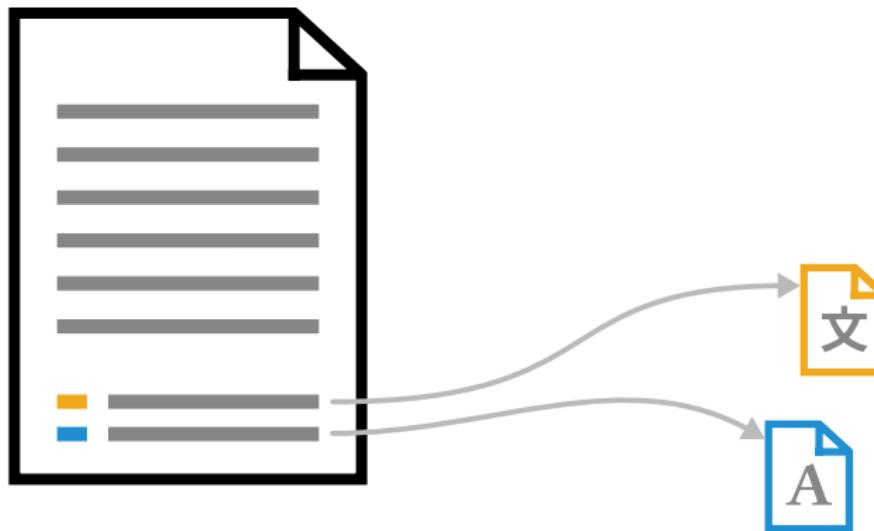
Citations & Non-English
ooo

Artifact Parameters
oooooooooooooooo

Conclusion
oooooo

References

Maps of Science



Motivation
ooooo

Background
○●○○

Outline
○○

Corpus
oooooooooo

Citations & Non-English
ooo

Artifact Parameters
oooooooooooooooo

Conclusion
oooooo

References

Maps of Science



Motivation
ooooo

Background
○●○○

Outline
○○

Corpus
oooooooooo

Citations & Non-English
ooo

Artifact Parameters
oooooooooooooooo

Conclusion
oooooo

References

Maps of Science



Motivation
ooooo

Background
○●○○

Outline
○○

Corpus
oooooooooo

Citations & Non-English
ooo

Artifact Parameters
oooooooooooooooo

Conclusion
oooooo

References

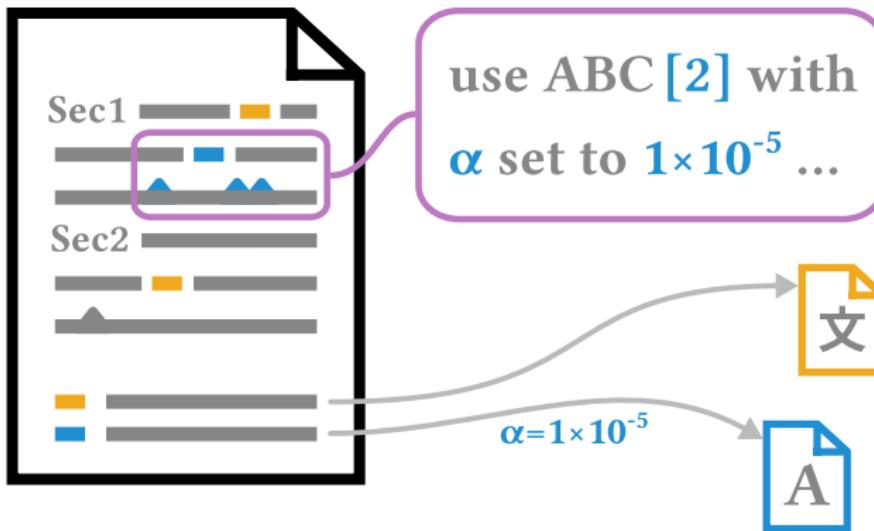
Maps of Science



use ABC [2] with
 α set to 1×10^{-5} ...



Maps of Science



Motivation
ooooo

Background
○●○○

Outline
○○

Corpus
oooooooooo

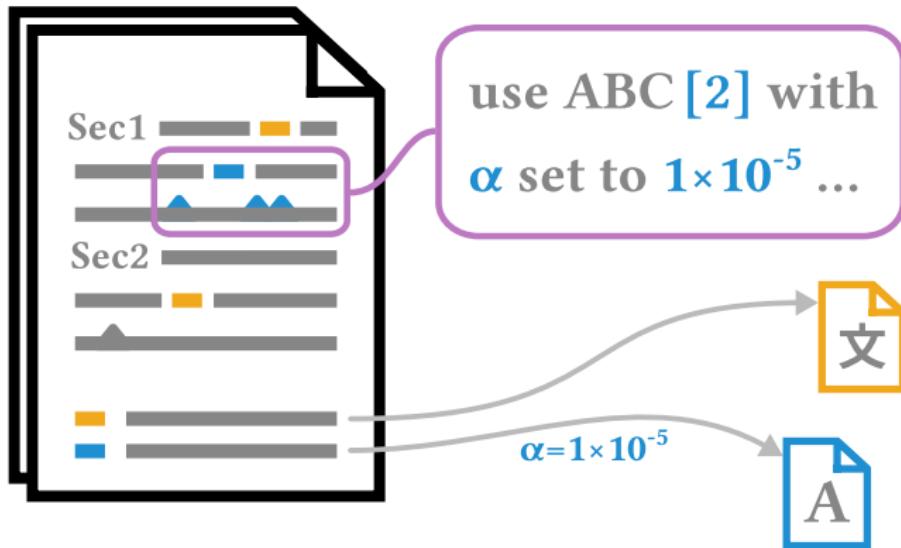
Citations & Non-English
ooo

Artifact Parameters
oooooooooooooooo

Conclusion
oooooo

References

Maps of Science



Motivation
ooooo

Background
●●○○

Outline
○○

Corpus
oooooooooo

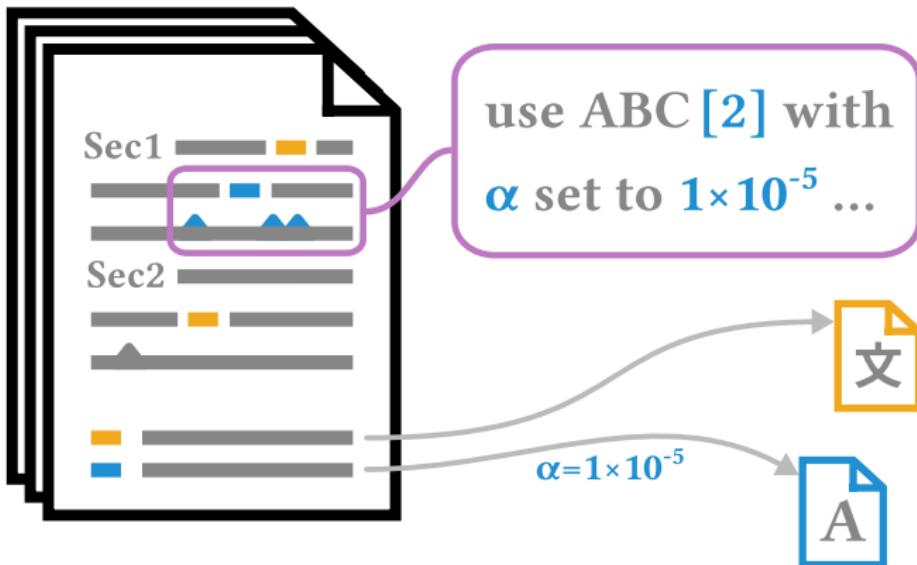
Citations & Non-English
ooo

Artifact Parameters
oooooooooooooooo

Conclusion
oooooo

References

Maps of Science



Motivation
ooooo

Background
●●○○

Outline
○○

Corpus
oooooooooo

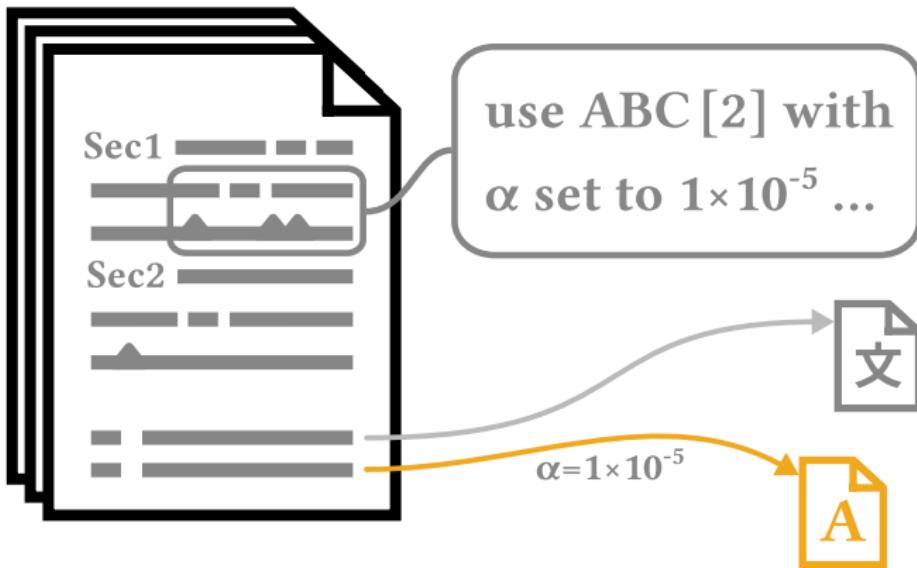
Citations & Non-English
ooo

Artifact Parameters
oooooooooooooooo

Conclusion
oooooo

References

Research Gap: Citation Network



Motivation
ooooo

Background
○○●○

Outline
○○

Corpus
oooooooooo

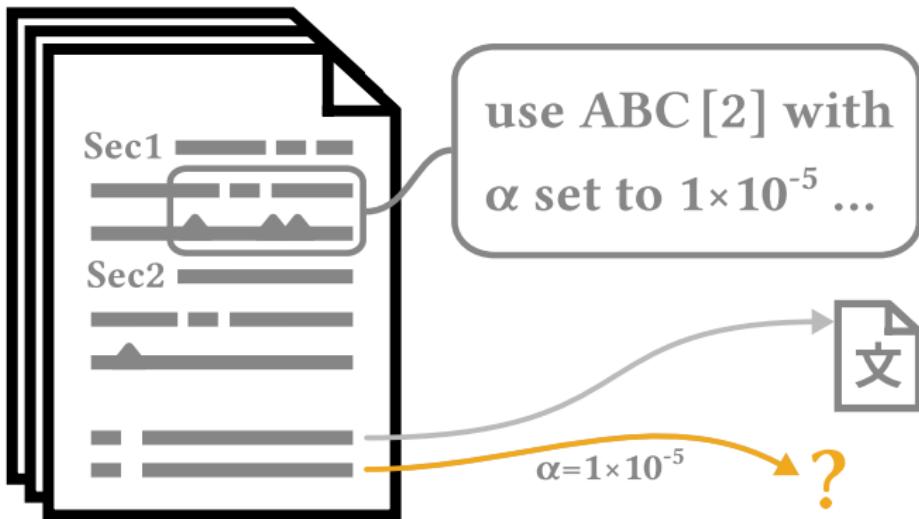
Citations & Non-English
ooo

Artifact Parameters
oooooooooooooooo

Conclusion
oooooo

References

Research Gap: Citation Network



Motivation
ooooo

Background
○○●○

Outline
○○

Corpus
oooooooooo

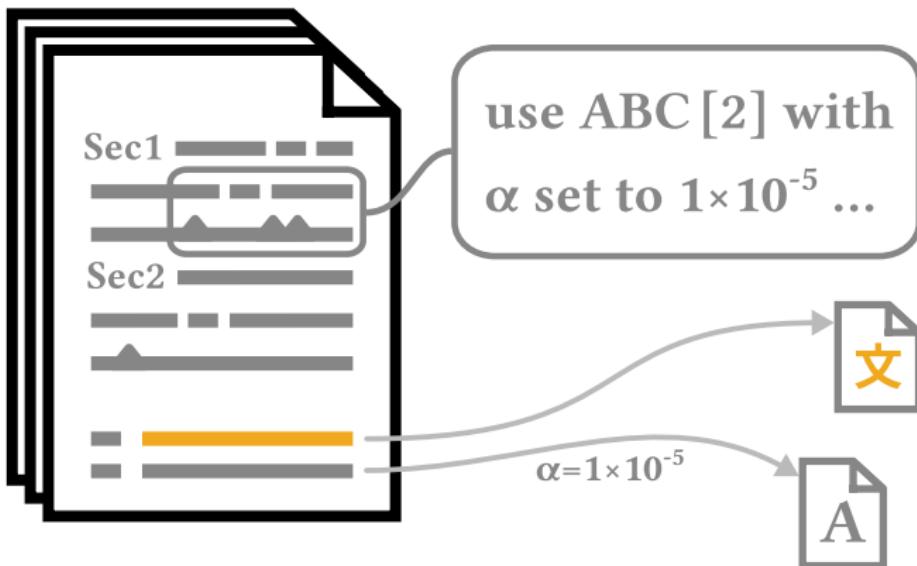
Citations & Non-English
ooo

Artifact Parameters
oooooooooooooooo

Conclusion
oooooo

References

Research Gap: Non-English Documents



Motivation
ooooo

Background
○○●○

Outline
○○

Corpus
oooooooooo

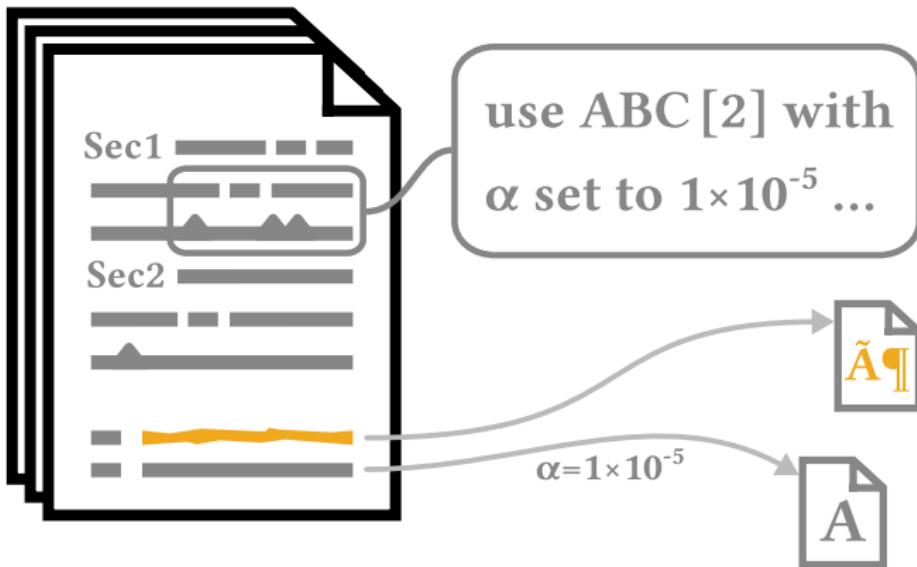
Citations & Non-English
ooo

Artifact Parameters
oooooooooooooooo

Conclusion
oooooo

References

Research Gap: Non-English Documents



Motivation
ooooo

Background
○○●○

Outline
○○

Corpus
oooooooooo

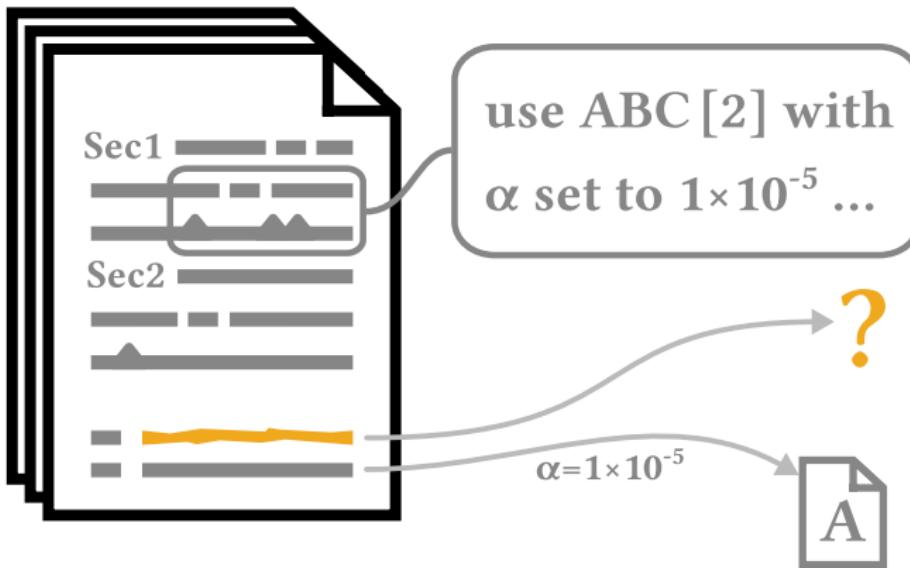
Citations & Non-English
ooo

Artifact Parameters
oooooooooooooooo

Conclusion
oooooo

References

Research Gap: Non-English Documents



Motivation
ooooo

Background
ooo●o

Outline
oo

Corpus
oooooooooooo

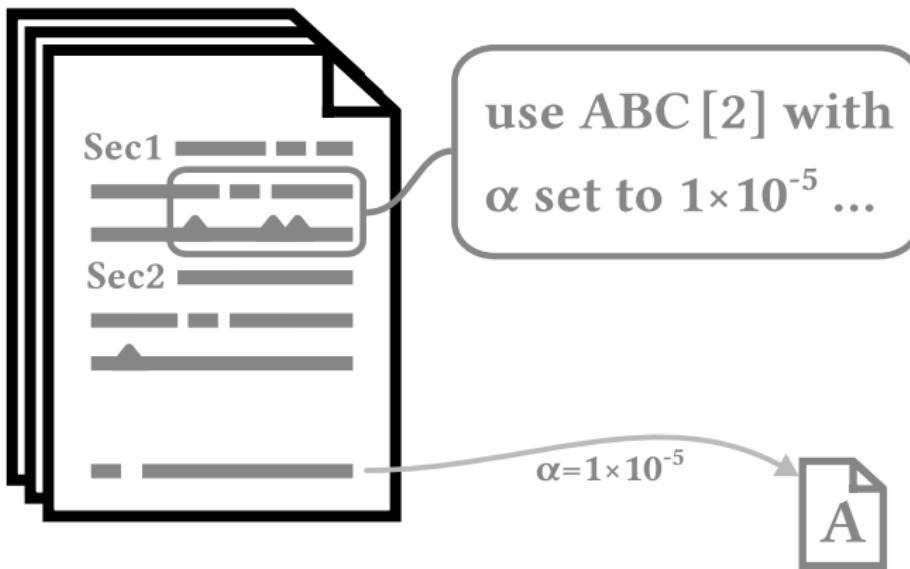
Citations & Non-English
ooo

Artifact Parameters
oooooooooooooooooooo

Conclusion
oooooo

References

Research Gap: Non-English Documents



Motivation
ooooo

Background
○○●○

Outline
○○

Corpus
oooooooo

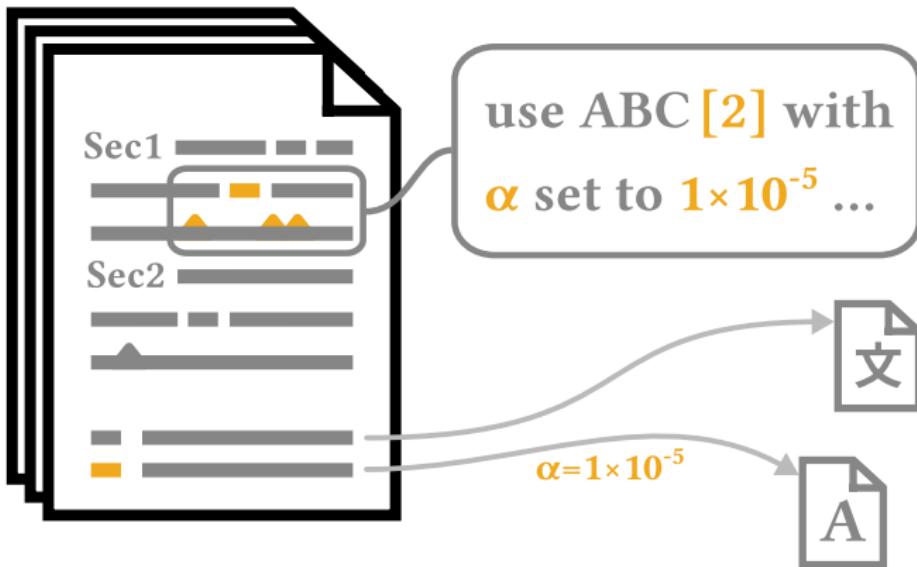
Citations & Non-English
ooo

Artifact Parameters
oooooooooooo

Conclusion
ooooo

References

Research Gap: Artifact Parameters



Motivation
ooooo

Background
○○●○

Outline
○○

Corpus
oooooooooo

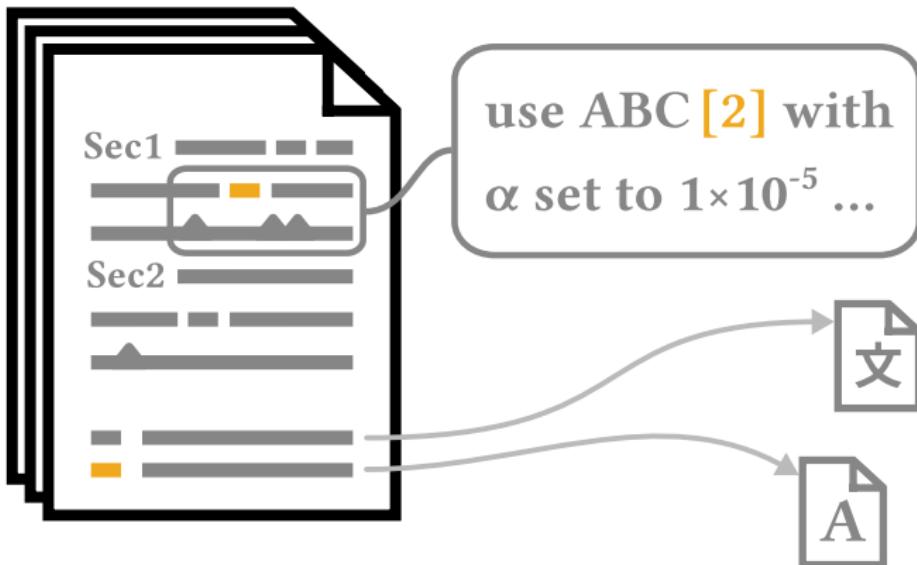
Citations & Non-English
ooo

Artifact Parameters
oooooooooooooooo

Conclusion
oooooo

References

Research Gap: Artifact Parameters



Motivation
ooooo

Background
○○●○

Outline
○○

Corpus
oooooooooo

Citations & Non-English
ooo

Artifact Parameters
oooooooooooooooo

Conclusion
oooooo

References

Data Mining and Information Extraction Methods

for Large-Scale High Quality Representations of Scientific Publications



Research Objective



Research Objective

Research Gap



Research Objective

Base

RT1: Corpus

Research Gap

RT2: Citations

RT3: Non-English

RT4: Artifacts



Research Objective



Data Quality

Base

RT1: Corpus

Research Gap

RT2: Citations

RT3: Non-English

RT4: Artifacts



Research Objective



Data Quality



Base

RT1: Corpus

Research Gap

RT2: Citations

RT3: Non-English

RT4: Artifacts



Research Objective



Data Quality

Base

Research Gap

RT1: Corpus

RT2: Citations

RT3: Non-English

RT4: Artifacts



✓ | X



✓ | X



✓ | X



✓ | X



✓ | X

✓ | X

✓ | X

✓ | X

✓ | X

✓ | X

✓ | X

✓ | X

✓ | X

✓ | X

✓ | X

✓ | X

✓ | X

✓ | X

✓ | X

✓ | X



Research Objective



Data Quality



Base

RT1: Corpus

- ① conversion and linking pipeline
- ② extensive corpus analysis

[Scientometrics'20]

RT2: Citations

- ① improved corpus pipeline
- ② blocking method for reference linking

[ULITE'22]
[JCDL'23]

Research Gap

RT3: Non-English

- ① extraction method for cross-lingual cit.
- ② extensive citation analysis

[ICADL'20]
[IJDL'22]

RT4: Artifacts

- ① novel IE task and data set
- ② IE model development and eval.

[ECIR'24]



Research Objective



Data Quality



Base

RT1: Corpus

- ① conversion and linking pipeline
- ② extensive corpus analysis

[Scientometrics'20]

Research Gap

RT2: Citations

- ① improved corpus pipeline
- ② blocking method for reference linking

[ULITE'22]
[JCDL'23]

RT3: Non-English

- ① extraction method for cross-lingual cit.
- ② extensive citation analysis

[ICADL'20]
[IJDL'22]

RT4: Artifacts

- ① novel IE task and data set
- ② IE model development and eval.

[ECIR'24]

Outline

■ Corpus

- Challenges
- Methods
- Resulting Corpus

■ Citation Network

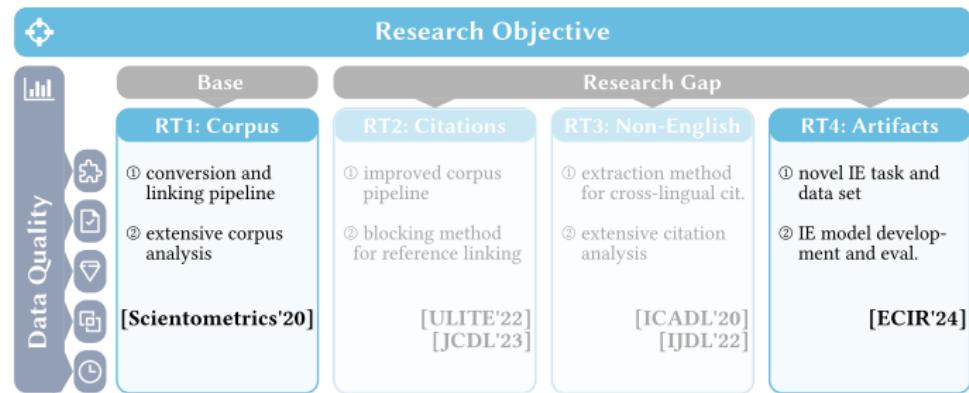
■ Non-English Documents

■ Artifact Parameters

- Task Definition
- Methods
- Results

■ Conclusion

- Contributions
- Impact



Motivation
ooooo

Background
oooo

Outline
○●

Corpus
oooooooooo

Citations & Non-English
ooo

Artifact Parameters
oooooooooooooooo

Conclusion
oooooo

References

Corpus

unarXive

Corpus - Digest

■ Research Task

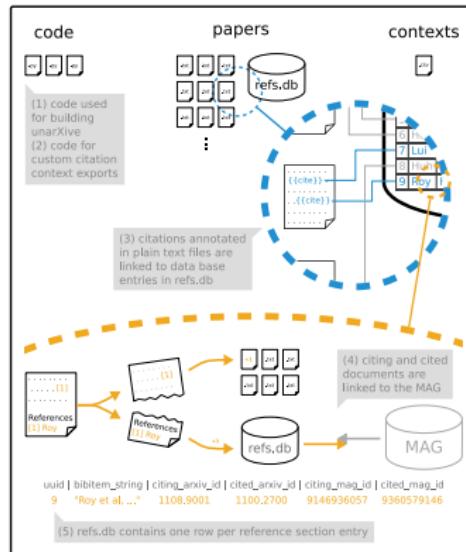
Develop a method for creating large-scale, high quality “base corpus”

■ Method

- L^AT_EX as data source
- Joint handling of text + references
- Reference parsing + linking to metadata records

■ Results

- Corpus among 3 largest full-text corpora
- More extensive, complete, less noisy
- State of the art citation network



Scientometrics'20 [1]

Motivation
ooooo

Background
oooo

Outline
oo

Corpus
○●oooooooo

Citations & Non-English
ooo

Artifact Parameters
oooooooooooo

Conclusion
oooooo

References

Corpus - Challenges

■ General

- Volume ($\sim 10^6$ docs, $\sim 10^7$ refs)
- Bridging visual medium and text information

■ Parsing

- Parser efficiency
- Typesetting info \neq semantic info
- \LaTeX is powerful and people are creative

■ Reference linking

- Choice of target set
- Parsing (bbl, not bib)
- Variance and information sparsity

```
\begin{document}
\newcommand{\nc}{\newcommand}
\nc{\be}{\begin{equation}}
\nc{\ee}{\end{equation}}
\nc{\bib}{\bibitem}
```

[10] I. Bonalde et al., Phys. Rev. Lett. **85**, 4775 (2000).

[25] Bonalde I, Yanoff B D, Salomon M B, Van Harlingen D J, Chia E M E, Mao Z Q and Maeno Y 2000
Phys. Rev. Lett. **85** 4775

[4] Jaume, S.C. and Sykes, L.R., Pure and Applied Geophysics **155**, 279-305.

Jaume, S.C. and L.R. Sykes, Evolving Towards a Critical Point: A Review of Accelerating Seismic
Moment/Energy Release Prior to Large and Great Earthquakes, Pure Appl. Geophys., 155, 279, 1999.

Motivation
ooooo

Background
oooo

Outline
oo

Corpus
oo●oooooooo

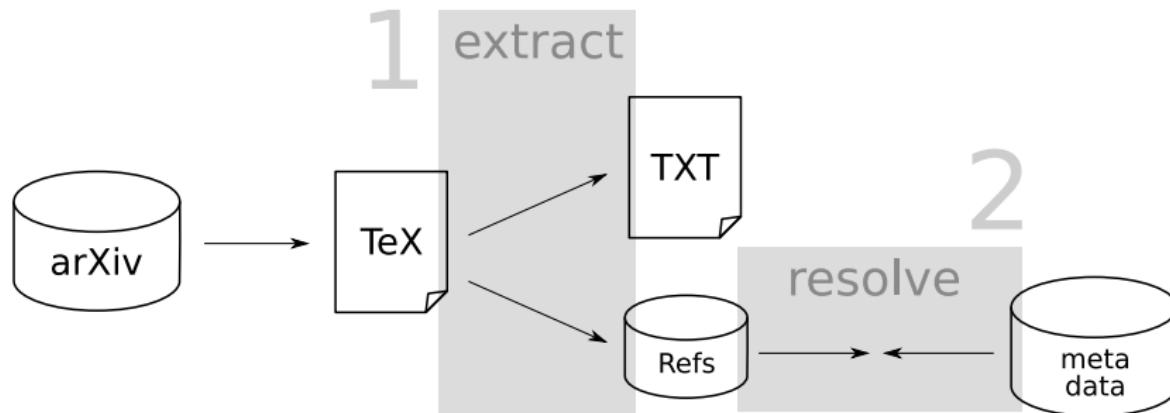
Citations & Non-English
ooo

Artifact Parameters
oooooooooooooooo

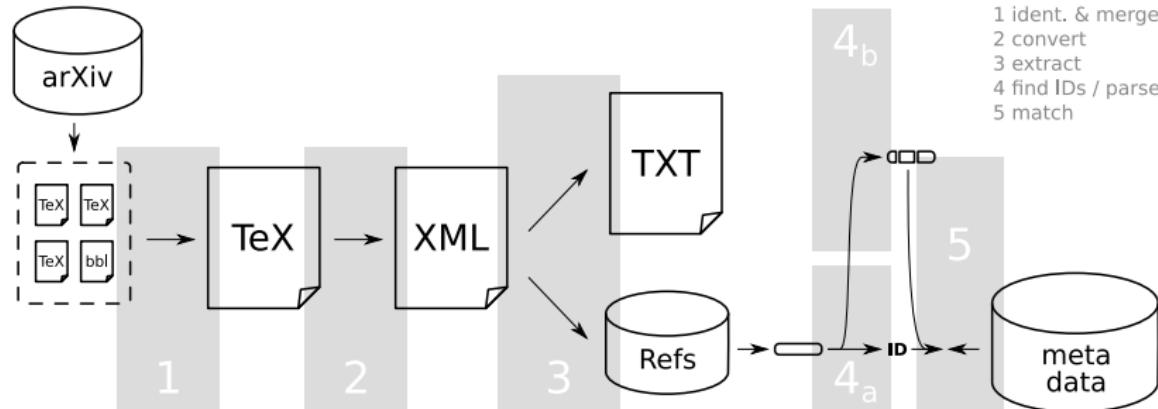
Conclusion
oooooo

References

Corpus - Methods



Corpus - Methods



Motivation
ooooo

Background
oooo

Outline
oo

Corpus
ooo●ooooo

Citations & Non-English
ooo

Artifact Parameters
oooooooooooooooo

Conclusion
oooooo

References

Corpus - Result

■ Size

- 1.2 M documents (2.7 M cited)
- 16 M references
- 29 M in-text citation markers

■ Scope

- 1991–2018 (current: 2022)
- physics (63%), maths (23%), CS (11%), other (3%)

■ Reference matching

- 53% by parsing + matching
- 28% by DOI
- 19% by arXiv ID

Motivation
ooooo

Background
oooo

Outline
oo

Corpus
oooo●oooo

Citations & Non-English
ooo

Artifact Parameters
oooooooooooo

Conclusion
oooooo

References

Corpus - Result

Data Set	# Docs	Cit. Markers	Linked	Link Succ.	Link Acc.
ACL-ARC [2]	11 k	×	×	n/a	n/a
ACL-AAN [3]	18 k	×	×	n/a	n/a
Scholarly Dataset 2 [4]	100 k	×	×	n/a	n/a
CiteSeerX [5] / RefSeer [6]	1 M	ambiguous	×	n/a	n/a
PMC OAS [7]	2.3 M	exact	mixed ^a	-	-
arXiv CS [8]	90 k	exact	✓	39.3%	-
unarXive [1]	1.2 M	exact	✓	42.6%	96%

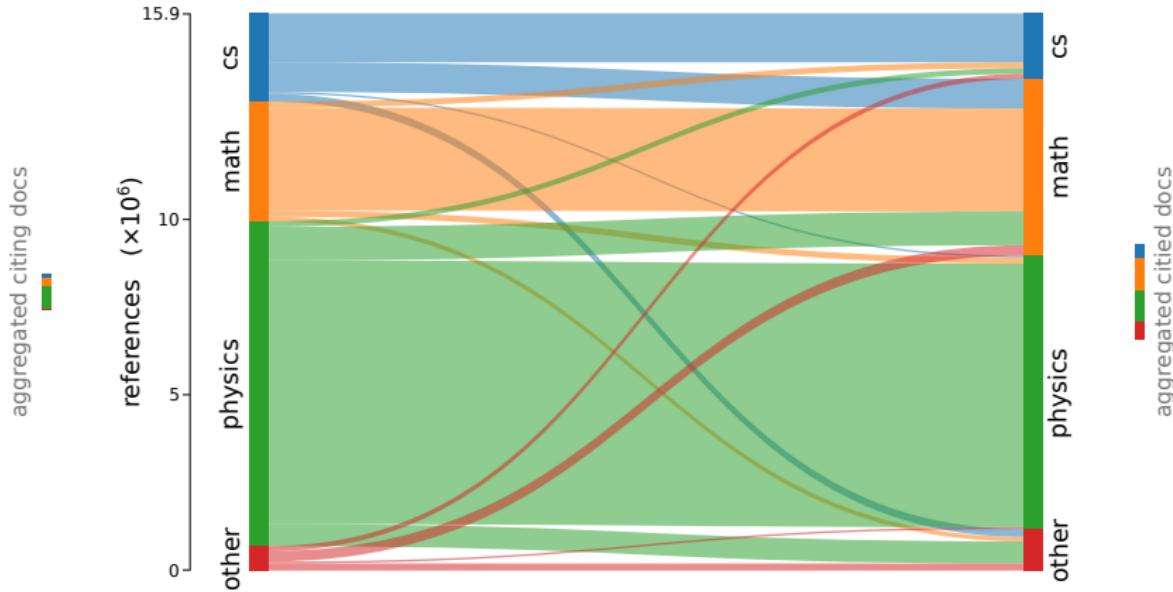
^a No citation network due to mixed set of IDs (PubMed, MEDLINE, DOI) [9].

Corpus - Result (2022)

Data Set	# Docs	Cit. Markers	Linked	Link Succ.	Link Acc.
arXMLiv [12]	1.6 M	exact	✗	n/a	n/a
PMC-OAS [7]	3.3 M	exact	mixed ^a	-	-
SciXGen [13]	205 k	exact	✓	41.6%	-
S2ORC (\LaTeX) [11]	1.5 M	exact	✓	31.1%	92%
unarXive 2022 [14]	1.9 M	exact	✓	44.4%	96%

^a No citation network due to mixed set of IDs (PubMed, MEDLINE, DOI) [9].

Corpus - Citation Flow



Motivation
ooooo

Background
oooo

Outline
oo

Corpus
oooooooo●○

Citations & Non-English
ooo

Artifact Parameters
oooooooooooooooo

Conclusion
ooooo

References

Corpus - Conclusion

■ Advancements

- More extensive
 - single domain vs multi domain
- More complete, high quality citation network
 - 13.3% increase in linked references (\leftrightarrow S2ORC_{LATEX})
 - 4% more accurate reference links (\leftrightarrow S2ORC_{LATEX})
- Less noise due to LATEX as source
- Novel types of analyses possible

■ Foundation for further studies

- →  RT1 ✓

Motivation
ooooo

Background
oooo

Outline
oo

Corpus
oooooooo●

Citations & Non-English
ooo

Artifact Parameters
oooooooooooooooo

Conclusion
ooooo

References

Citations & Non-English

Citation Network - Digest

■ Research Task

Develop a method linking references more successfully without compromising accuracy

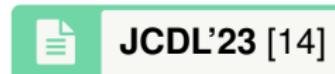
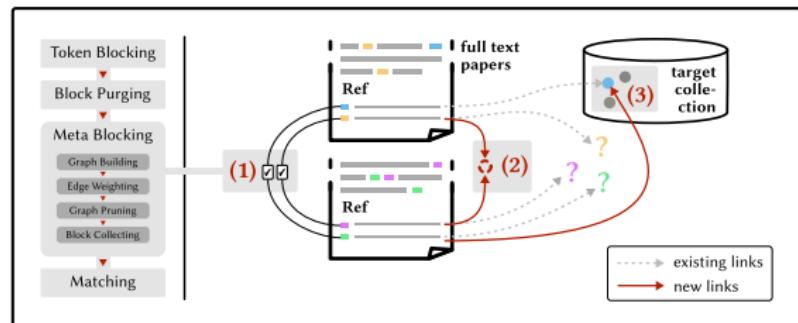
■ Method

- Use unarXive data
- Improved reference linking pipeline
(parser & target set choice, heuristics)
- Blocking & matching within set of references

■ Results

- +2% in base matching success (SOTA)
- Manifold increase in bibl. couplings

■ →  RT2✓



Motivation
ooooo

Background
oooo

Outline
oo

Corpus
oooooooooooo

Citations & Non-English
○●○

Artifact Parameters
oooooooooooooooooooo

Conclusion
ooooooo

References

Non-English Documents - Digest

■ Research Task

Develop an approach to include non-English publications into large-scale scholarly data

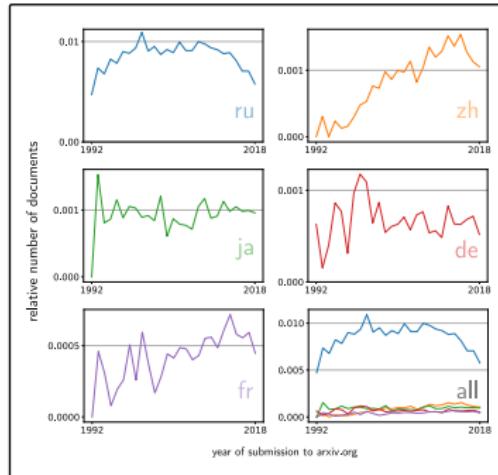
■ Method

- Use unarXive data
- Identify cross-lingual citations by reference strings
- Temporal and geographic analyses

■ Results

- Reliable method for identification
- Largest study so far ($<1k \rightarrow >1M$)
- Identification of trends and challenges

■ → RT3✓



Motivation
ooooo

Background
oooo

Outline
oo

Corpus
oooooooooooo

Citations & Non-English
oo●

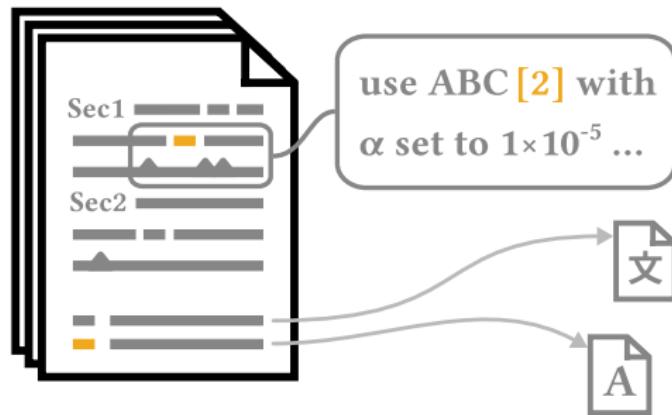
Artifact Parameters
oooooooooooooooooooo

Conclusion
oooooo

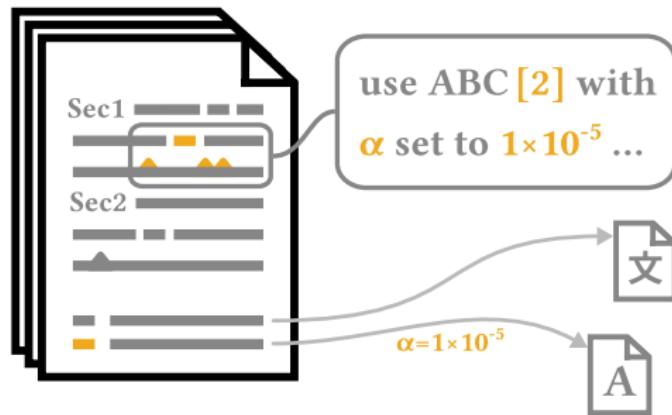
References

Artifact Parameters

Artifact Parameters

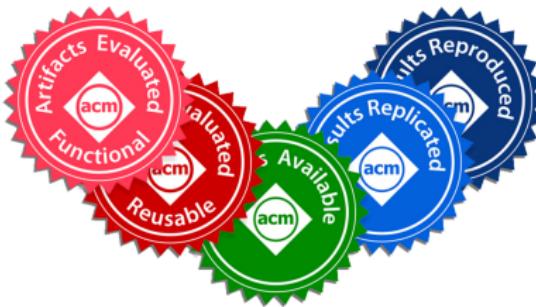


Artifact Parameters



Artifact Parameters - Motivation

The image shows two side-by-side screenshots of web interfaces. On the left, the Databricks Research homepage features a navigation bar with 'Search', 'Browse State-of-the-Art', 'Datasets', 'Methods', and 'More'. Below this are filters for 'Tip', 'New', and 'Greatest'. A section titled 'Trending Research' highlights 'Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction' by FoundationVision/VaR, with a rating of 1.386 stars and 5.5k stars/reviews. The right screenshot shows a 'Dataset Search' interface with a Google logo at the top. It includes a search bar, a 'Search for Datasets' button, and a message encouraging users to try 'coronavirus covid-19' or 'water quality site:canada.ca'. A 'Learn more' link is also present.



Motivation

oooooo

Background

Outline
oo

Corpus

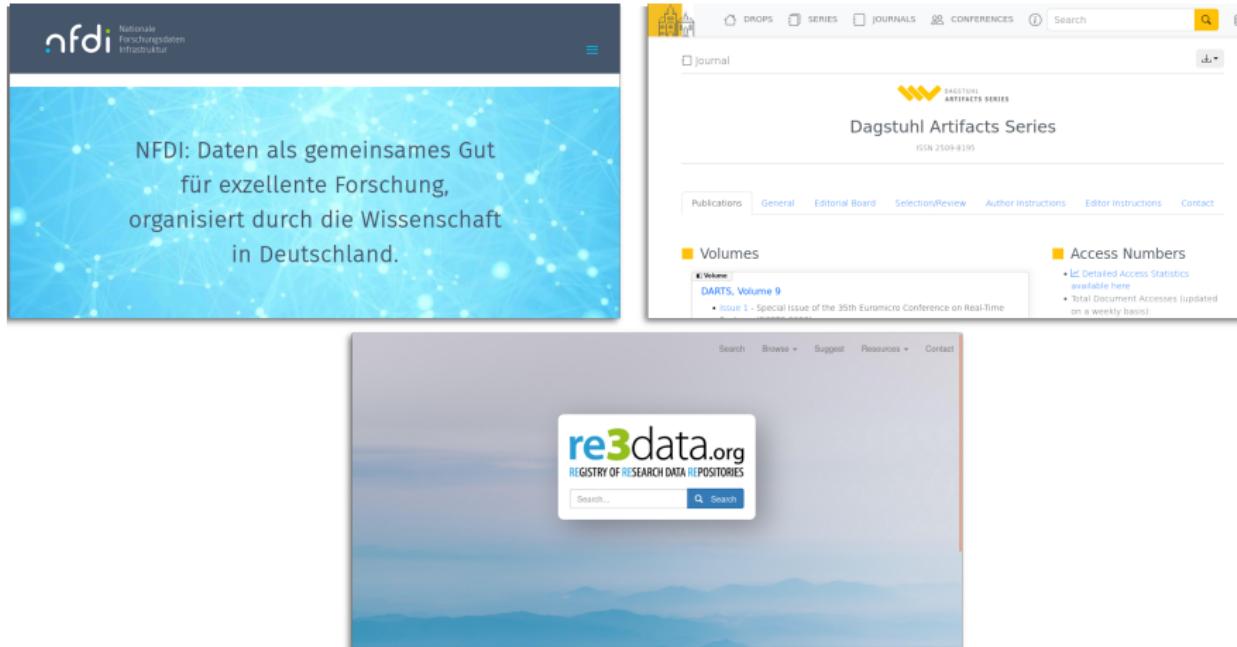
Citations & Non-English ooo

Artifact Parameters

Conclusion

References

Artifact Parameters - Motivation



Motivation
ooooo

Background
oooo

Outline
oo

Corpus
oooooooo

Citations & Non-English
ooo

Artifact Parameters
oo●oooooooooooo

Conclusion
ooooo

References

Artifact Parameters - Motivation

NLP field increasing focus on **data** and its **algorithmic processing**

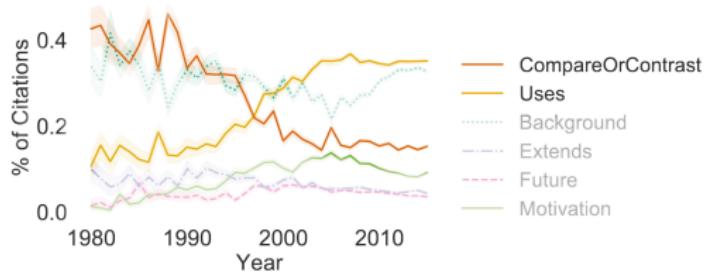


Figure 5: Changes in the average citation frame in ACL papers reveals a continued decline in the percentage of COMPARISON OR CONTRAST and increase in USES citations. The increase in BACKGROUND citations circa 2010 marks the start of the era of unlimited references in ACL conferences. Shaded regions show bootstrapped 95% confidence intervals.

Jurgens, D. et al. Measuring the Evolution of a Scientific Field through Citation Frames. *Transactions of the Association for Computational Linguistics* 6, 391–406 (2018)

Artifact Parameters - Digest

■ Research Task

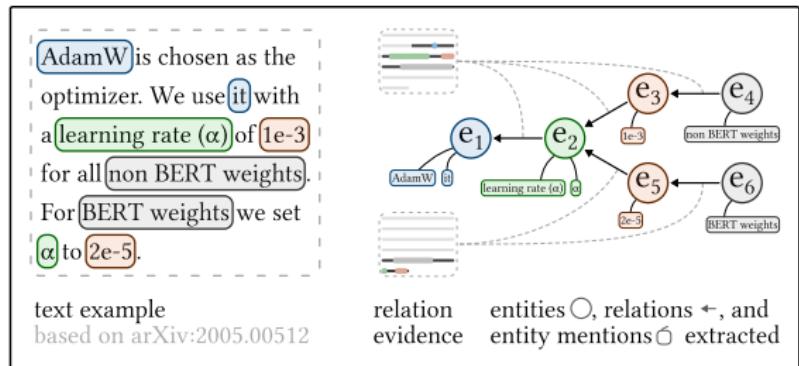
Develop a method to extract fine-grained research artifacts information

■ Method

- Task formalization
- Data annotation (data from unarXive)
- Two lines of approaches
 - BERT based model approach
 - LLM based approach

■ Results

- Novel task, novel data
- Improvements over SOTA baselines
- Methods applicable to large data sets



ECIR'24 [20]

Motivation
ooooo

Background
oooo

Outline
oo

Corpus
oooooooo

Citations & Non-English
ooo

Artifact Parameters
oooooooooooo

Conclusion
ooooo

References

Artifact Parameters - Task

AdamW is chosen as the optimizer. We use it with a learning rate (α) of 1e-3 for all non BERT weights. For BERT weights we set α to 2e-5.

text example
based on arXiv:2005.00512

Motivation
ooooo

Background
oooo

Outline
oo

Corpus
oooooooo

Citations & Non-English
ooo

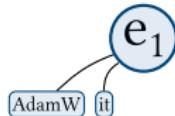
Artifact Parameters
oooo●oooooooo

Conclusion
ooooo

References

Artifact Parameters - Task

AdamW is chosen as the optimizer. We use it with a learning rate (α) of 1e-3 for all non BERT weights. For BERT weights we set α to 2e-5.

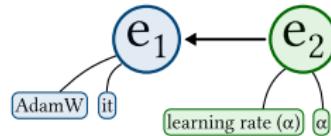


text example
based on arXiv:2005.00512

entity type
artifact

Artifact Parameters - Task

AdamW is chosen as the optimizer. We use it with a learning rate (α) of 1e-3 for all non BERT weights. For BERT weights we set α to 2e-5.



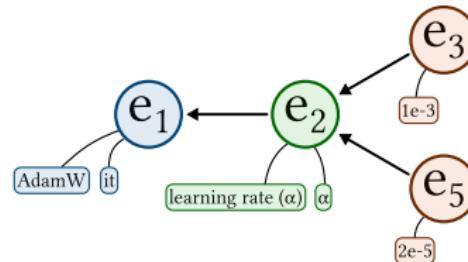
text example
based on arXiv:2005.00512

entity type
artifact parameter

Artifact Parameters - Task

AdamW is chosen as the optimizer. We use it with a learning rate (α) of $1e-3$ for all non BERT weights. For BERT weights we set α to $2e-5$.

text example
based on arXiv:2005.00512

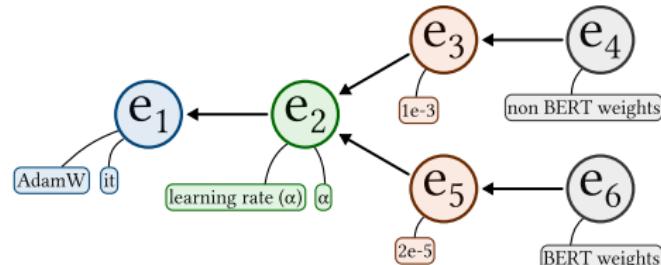


entity type
artifact parameter value

Artifact Parameters - Task

AdamW is chosen as the optimizer. We use it with a learning rate (α) of $1e-3$ for all non BERT weights. For BERT weights we set α to $2e-5$.

text example
based on arXiv:2005.00512



entity type
artifact parameter value context

Artifact Parameters - Related Work: Fine-tuned models

- SciERC dataset
 - SCIEE [21]
 - PL-Marker [22]
 - → **entity type overlap**
- [...]
- SemEval 2021 (measurements)
 - LIORI [23]
 - → **utilize mention pattern regularities**

We evaluate our model on the task of **question answering** using

Section : Dataset

SQuAD is a **machine comprehension** dataset on a large set of **Wikipedia articles** , Two metrics are used to evaluate models: **Exact Match (EM)** and a softer metric , **F1 score**

Section: Model Details .

... Each paragraph and question are tokenized by a regular - expression - based word tokenizer (**PTB Tokenizer**) and fed into the model .
....

Section : Results .

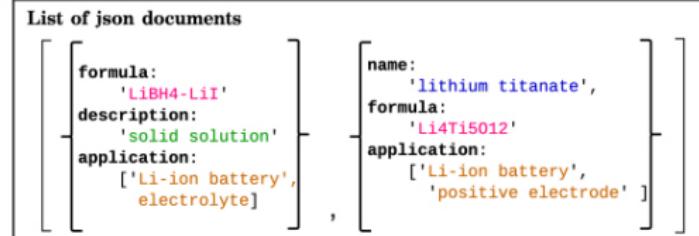
The results of our model and competing approaches on the hidden test are summarized in Table [reference]. **BiDAF (ensemble)** achieves an **EM** score of 73.3 and an **F1** score of 81.1 , outperforming all previous approaches .

Artifact Parameters - Related Work: LLMs

- Medical science [24]
 - singular values
 - lists
- Material science [25–27]
 - singular values
 - lists
 - hierarchical [27] (see right)
→ **data serialization format**

Note: all of the above evaluate
on **GPT models only**.

Document: The charge and discharge performance of an all-solid-state lithium battery with the LiBH4-LiI solid solution as an electrolyte is reported. Lithium titanate ($\text{Li}_4\text{Ti}_5\text{O}_{12}$) was used as the positive electrode and...



Dunn, A. et al. *Structured information extraction from complex scientific text with fine-tuned large language models*. Dec. 2022. doi:10.48550/arXiv.2212.05238

Motivation
ooooo

Background
oooo

Outline
oo

Corpus
oooooooo

Citations & Non-English
ooo

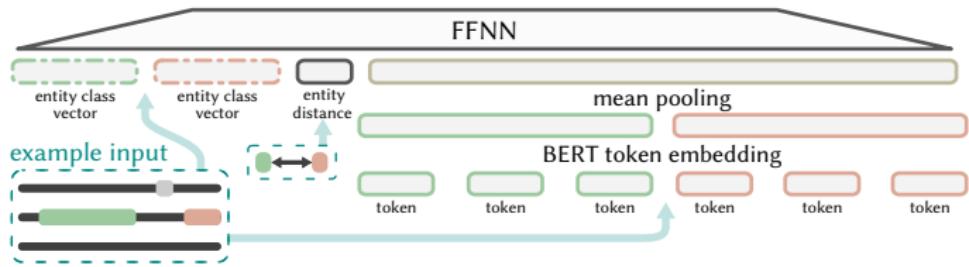
Artifact Parameters
oooooooo●oooooooo

Conclusion
ooooo

References

Approach: Fine-tuned models

- Based on PL-Marker (SciERC SOTA)
- (N)ER: used as is
- RE: new module, utilizing
 - Entity class embeddings
 - Entity distance



Motivation
ooooo

Background
oooo

Outline
oo

Corpus
oooooooo

Citations & Non-English
ooo

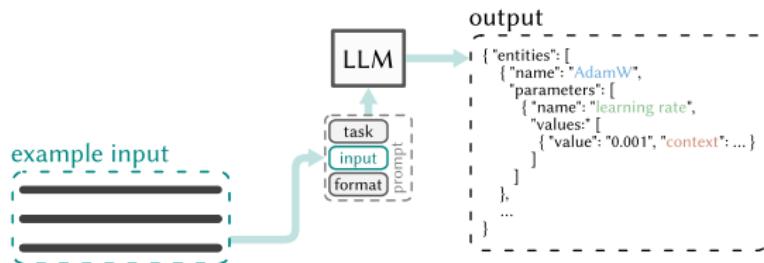
Artifact Parameters
oooooooo●oooooooo

Conclusion
ooooo

References

Approach: LLMs

- Data serialization format
 - JSON → YAML
- Compare 6 models
- Base prompt + tuning for each
- Zero-shot: all
- Few-shot: only Vicuna_{16k}



Approach: LLMs

- Data serialization format
 - JSON → YAML
- Compare 6 models
- Base prompt + tuning for each
- Zero-shot: all
- Few-shot: only Vicuna_{16k}

Model	Size
WizardLM [28]	13 B
Vicuna _{4k} [29]	13 B
Vicuna _{16k} [29]	13 B
Falcon [30]	40 B
GALACTICA [31]	120 B
GPT-3.5 [32]	175 B

Motivation
ooooo

Background
oooo

Outline
oo

Corpus
oooooooo

Citations & Non-English
ooo

Artifact Parameters
oooooooo●oooo

Conclusion
ooooo

References

Artifact Parameters - Data

- No existing data sets
→ **use unarXive** ML/CV/CL/DL
 - Annotation approach
 - Paragraph level, whole papers
 - **Data Set Size**
 - 444 paragraphs
 - 1,971 entities
 - 614 relations
 - IAA
 - 0.867 for entities
 - 0.737 for relations
 - → **high quality**

Motivation

Background

Outline

Corpus

Citations & Non-English ooo

Artifact Parameters

Conclusion

References

Artifact Parameters - Experiments: Fine-tuned models

■ Setting

- 5-fold cross-validation
- Stratified sampling

■ Results (F_1 [%])

- ER: 78.0
- RE: 9.9 → 38.8

■ Analysis

- Parameter: low performance
- Contexts: not predicted

Motivation
ooooo

Background
oooo

Outline
oo

Corpus
oooooooo

Citations & Non-English
ooo

Artifact Parameters
oooooooo●ooo

Conclusion
ooooo

References

Artifact Parameters - Experiments: Fine-tuned models

■ Setting

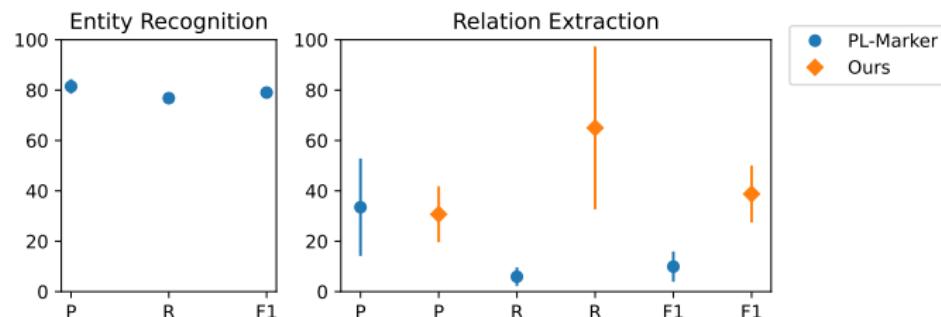
- 5-fold cross-validation
- Stratified sampling

■ Results (F_1 [%])

- ER: 78.0
- RE: 9.9 → 38.8

■ Analysis

- Parameter: low performance
- Contexts: not predicted



Artifact Parameters - Experiments: Fine-tuned models

■ Setting

- 5-fold cross-validation
- Stratified sampling

■ Results (F_1 [%])

- ER: 78.0
- RE: 9.9 → 38.8

■ Analysis

- Parameter: low performance
- Contexts: not predicted

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

Used	P [%]	R [%]	F_1 [%]
¬CD	15.5	8.8	11.1
T¬D	16.6	29.8	19.6
TC¬	26.5	65.0	35.5
TCD	30.7	65.0	38.8

Artifact Parameters - Experiments: LLMs

■ Setting

- Zero-/5-shot
- Compare JSON w/ YAML variant

■ Results (best ER / RE F₁ [%])

- Zero-shot: 37.4 / 7.8
- 5-shot: 44.0 / 6.1

■ Analysis

- Very low RE performance
- YAML: avg. +5% in ER
- 5-shot: +27% ER, +6% RE
- Format adherence
+ entity hallucinations

Motivation
ooooo

Background
oooo

Outline
oo

Corpus
oooooooo

Citations & Non-English
ooo

Artifact Parameters
oooooooooooo●ooo

Conclusion
ooooo

References

Artifact Parameters - Experiments: LLMs

■ Setting

- Zero-/5-shot
- Compare JSON w/ YAML variant

■ Results (best ER / RE F_1 [%])

- Zero-shot: 37.4 / 7.8
- 5-shot: 44.0 / 6.1

■ Analysis

- Very low RE performance
- YAML: avg. +5% in ER
- 5-shot: +27% ER, +6% RE
- Format adherence
+ entity hallucinations

Zero-shot		ER	RE
Model	Output	F_1 [%]	F_1 [%]
WizardLM	JSON	8.6	0.1
	YAML	15.3 <small>△+6.7</small>	0.1 <small>△+0.0</small>
Vicuna _{4k}	JSON	11.5	1.2
	YAML	22.3 <small>△+10.8</small>	0.1 <small>△-1.1</small>
Falcon	JSON	10.2	0.0
	YAML	19.8 <small>△+9.6</small>	0.0 <small>△+0.0</small>
GALACTICA	JSON	19.5	0.3
	YAML	21.1 <small>△+1.6</small>	0.1 <small>△-0.2</small>
GPT-3.5	JSON	33.8	7.2
	YAML	37.4 <small>△+3.6</small>	7.8 <small>△+0.6</small>
5-shot		ER	RE
Vicuna _{16k}	JSON	39.6	1.3
	YAML	44.0 <small>△+0.4</small>	6.1 <small>△+4.8</small>

Motivation
ooooo

Background
oooo

Outline
oo

Corpus
oooooooooooo

Citations & Non-English
ooo

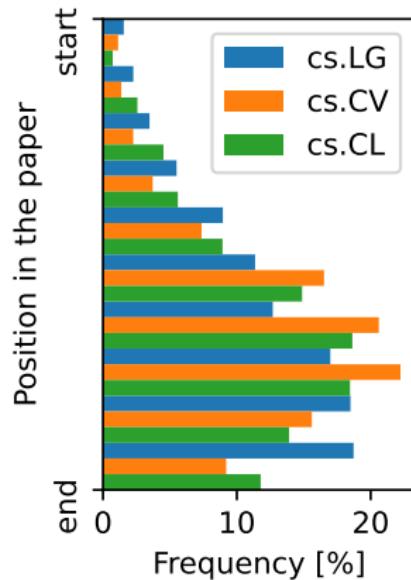
Artifact Parameters
oooooooooooo●ooo

Conclusion
oooooo

References

Artifact Parameters - Experiments: Application

- Apply best model (BERT based) on 15k paper sample
- Parameters information given in
 - 36% of ML papers
 - 42% of CV papers
 - 36% of CL papers
 - 7% of DL papers
- Distribution towards second half of paper across disciplines



Motivation
ooooo

Background
oooo

Outline
oo

Corpus
oooooooo

Citations & Non-English
ooo

Artifact Parameters
oooooooooooo●○

Conclusion
ooooo

References

Artifact Parameters - Conclusion

■ Advancements

- Novel, relevant task
(data scheme, annotation guidelines)
- High quality manually annotated data set
- Approaches based on BERT, LLMs
 - BERT model based approach
29% F_1 increase for RE
 - LLM approach
Avg. 5.5% F_1 increase for ER
(consistent across all used LLMs)
- Trained model applicable on large scale

■ →  RT4✓

Motivation
ooooo

Background
oooo

Outline
oo

Corpus
oooooooo

Citations & Non-English
ooo

Artifact Parameters
oooooooooooo●

Conclusion
ooooo

References

Conclusion



Research Objective



Research Objective

Research Gap



Research Objective

Base

RT1: Corpus

Research Gap

RT2: Citations

RT3: Non-English

RT4: Artifacts



Research Objective

Base

RT1: Corpus

- ① conversion and linking pipeline
- ② extensive corpus analysis

[Scientometrics'20]

RT2: Citations

- ① improved corpus pipeline
- ② blocking method for reference linking

[ULITE'22]
[JCDL'23]

Research Gap

RT3: Non-English

- ① extraction method for cross-lingual cit.
- ② extensive citation analysis

[ICADL'20]
[IJDL'22]

RT4: Artifacts

- ① novel IE task and data set
- ② IE model development and eval.

[ECIR'24]



Research Objective

Base

RT1: Corpus

- ① conversion and linking pipeline
- ② extensive corpus analysis

[Scientometrics'20]

RT2: Citations

- ① improved corpus pipeline
- ② blocking method for referent linking

[ULITE'22]
[JCDL'23]

Research Gap

RT3: Non-English

- ① extraction method for cross-lingual cit.
- ② extensive citation analysis

[ICADL'20]
[IJDL'22]

RT4: Artifacts

- ① novel IE task and data set
- ② IE model development and eval.

[ECIR'24]



Research Objective



Data Quality

Base

Research Gap

RT1: Corpus

RT2: Citations

RT3: Non-English

RT4: Artifacts



+

+

○

○



+

=

○

○



+

+

+

+



+

+

+

+



+

+

○

○



Research Objective ✓



Base

RT1: Corpus

- ① conversion and linking pipeline
- ② extensive corpus analysis

[Scientometrics'20]



Research Gap

RT2: Citations

- ① improved corpus pipeline
- ② blocking method for referential linking

[ULITE'22]
[JCDL'23]

RT3: Non-English

- ① extraction method for cross-lingual cit.
- ② extensive citation analysis

[ICADL'20]
[IJDL'22]

RT4: Artifacts

- ① novel IE task and data set
- ② IE model development and eval.

[ECIR'24]

Overall Conclusion

■ Research Objective

- Addressed research gaps in 3 areas of key importance
- All developed methods demonstrably applicable to large-scale data
- Comprehensive improvements of data quality

■ Impact on research community



Motivation
ooooo

Background
oooo

Outline
oo

Corpus
oooooooooooo

Citations & Non-English
ooo

Artifact Parameters
oooooooooooooooooooo

Conclusion
oo●ooo

References

Impact

Adoption by the research community.

■ Methodology

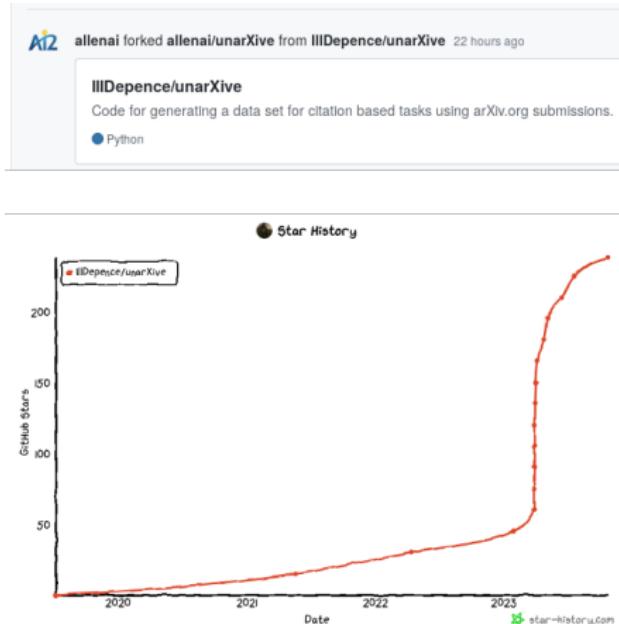
- Document Parsing Methodology [11]
 - Reference Matching Methodology [13]

■ Model dev/eval

- Citation Recommendation [33]
 - Document Retrieval [34]
 - Researcher Profile Embeddings [35]

■ Analyses

- Interdisciplinary Citations [36]
 - Citation Context Semantic Shifts [37]
 - Obliteration by Incorporation [38]



Big Picture

Compared to others

- **S2ORC** (Allen AI)
 - larger
 - less granular (maths), more noisy (PDF)
- **arXMLiv** (FAU KWARC)
 - more structured
 - no citation network
- **ORKG** (TIB)
 - more rigorously semantic
 - smaller (manual/semi-automated input)



Our work combines (1) **accurate, fine-granular** document representations,
(2) a **citation network**, and (3) applicability on a **large scale** due to being automated.

Motivation
ooooo

Background
oooo

Outline
oo

Corpus
oooooooo

Citations & Non-English
ooo

Artifact Parameters
oooooooooooo

Conclusion
oooo●o

References

Thank You ☺

Questions Discussion Comments

References I

1. Saier, T. & Färber, M. unarXive: A Large Scholarly Data Set with Publications' Full-Text, Annotated In-Text Citations, and Links to Metadata. *Scientometrics* **125**, 3085–3108. ISSN: 1588-2861 (Dec. 2020).
2. Bird, S. 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* (Marrakech, Morocco, 2008).
3. Radev, D. R., Muthukrishnan, P., Qazvinian, V. & Abu-Jbara, A. The ACL anthology network corpus. *Language Resources and Evaluation* **47**, 919–944 (2013).
4. Sugiyama, K. & Kan, M. A Comprehensive Evaluation of Scholarly Paper Recommendation Using Potential Citation Papers. *International Journal on Digital Libraries* **16**, 91–109 (2015).
5. Caragea, C. et al. *CiteSeer x : A Scholarly Big Dataset*. in *Proceedings of the 36th European Conference on IR Research* (Amsterdam, The Netherlands, 2014), 311–322.
6. Huang, W., Wu, Z., Liang, C., Mitra, P. & Giles, C. L. *A Neural Probabilistic Model for Context Based Citation Recommendation*. in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (AAAI Press, Austin, Texas, 2015), 2404–2410. ISBN: 0-262-51129-0.
7. Of Medicine, B. (N. L. *PMC Open Access Subset*. [Internet]. B. 2003 - [cited 2023 Feb 7]. <https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/>.
8. Färber, M., Thiemann, A. & Jatowt, A. *A High-Quality Gold Standard for Citation-based Tasks*. in *Proceedings of the 11th International Conference on Language Resources and Evaluation* (Miyazaki, Japan, 2018).

Motivation
ooooo

Background
oooo

Outline
oo

Corpus
oooooooooooo

Citations & Non-English
ooo

Artifact Parameters
oooooooooooooooooooo

Conclusion
oooooo

References

References II

9. Gipp, B., Meuschke, N. & Lipinski, M. *CITREC : An Evaluation Framework for Citation-Based Similarity Measures based on TREC Genomics and PubMed Central.* in *iConference 2015 Proceedings* (iSchools, 2015).
10. Pontika, N., Knoth, P., Cancellieri, M. & Pearce, S. Developing Infrastructure to Support Closer Collaboration of Aggregators with Open Repositories. *LIBER Quarterly* **25**, 172–188 (Apr. 2016).
11. Lo, K., Wang, L. L., Neumann, M., Kinney, R. & Weld, D. *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), 4969–4983.
12. Ginev, D. *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.
13. Chen, H., Takamura, H. & Nakayama, H. *SciXGen: A Scientific Paper Dataset for Context-Aware Text Generation.* in *Findings of the Association for Computational Linguistics: EMNLP 2021* (Nov. 2021), 1483–1492. doi:10.18653/v1/2021.findings-emnlp.128.
14. Saier, T., Krause, J. & Färber, M. *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, Los Alamitos, CA, USA, June 2023), 66–70. doi:10.1109/JCDL57899.2023.00020.
15. Brown, L. D., Cai, T. T. & DasGupta, A. Interval Estimation for a Binomial Proportion. *Statistical Science* **16**, 101–133 (2001).
16. Saier, T., Luan, M. & Färber, M. *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).

Motivation
ooooo

Background
oooo

Outline
oo

Corpus
oooooooooooo

Citations & Non-English
ooo

Artifact Parameters
oooooooooooooooooooo

Conclusion
oooooo

References

References III

17. Saier, T. & Färber, M. *A Large-Scale Analysis of Cross-lingual Citations in English Papers*. in *Digital Libraries at Times of Massive Societal Transition* (Springer International Publishing, 2020), 122–138. ISBN: 978-3-030-64452-9. doi:10.1007/978-3-030-64452-9_11.
18. Saier, T., Färber, M. & Tsereteli, T. Cross-Lingual Citations in English Papers: A Large-Scale Analysis of Prevalence, Formation, and Ramifications. en. *International Journal on Digital Libraries* 23, 179–195. ISSN: 1432-1300 (June 2022).
19. Jurgens, D., Kumar, S., Hoover, R., McFarland, D. & Jurafsky, D. Measuring the Evolution of a Scientific Field through Citation Frames. *Transactions of the Association for Computational Linguistics* 6, 391–406 (2018).
20. Saier, T., Ohta, M., Asakura, T. & Färber, M. *HyperPIE: Hyperparameter Information Extraction from Scientific Publications*. in *Advances in Information Retrieval 14609* (Springer Nature Switzerland, Mar. 2024), 254–269. ISBN: 978-3-031-56060-6. doi:10.1007/978-3-031-56060-6_17.
21. Luan, Y., He, L., Ostendorf, M. & Hajishirzi, H. *Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction*. in *Proc. Conf. Empirical Methods Natural Language Process. (EMNLP)* (2018).
22. Ye, D., Lin, Y., Li, P. & Sun, M. *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), 4904–4917. doi:10.18653/v1/2022.acl-long.337.
23. Davletov, A., Gordeev, D., Arefyev, N. & Davletov, E. *LIORI at SemEval-2021 Task 8: Ask Transformer for measurements*. in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)* (Aug. 2021), 1249–1254. doi:10.18653/v1/2021.semeval-1.178.

Motivation
ooooo

Background
oooo

Outline
oo

Corpus
oooooooooooo

Citations & Non-English
ooo

Artifact Parameters
oooooooooooooooooooo

Conclusion
oooooo

References

References IV

24. Agrawal, M., Hegselmann, S., Lang, H., Kim, Y. & Sontag, D. *Large language models are few-shot clinical information extractors.* in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (Dec. 2022), 1998–2022.
25. Xie, T. et al. *Large Language Models as Master Key: Unlocking the Secrets of Materials Science with GPT.* Apr. 2023. doi:10.48550/arXiv.2304.02213.
26. Polak, M. P. & Morgan, D. *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.
27. Dunn, A. et al. *Structured information extraction from complex scientific text with fine-tuned large language models.* Dec. 2022. doi:10.48550/arXiv.2212.05238.
28. Xu, C. et al. *WizardLM: Empowering Large Language Models to Follow Complex Instructions.* 2023.
29. Chiang, W.-L. et al. *Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.* Mar. 2023. <https://lmsys.org/blog/2023-03-30-vicuna/>.
30. Almazrouei, E. et al. *Falcon-40B: an open large language model with state-of-the-art performance.* 2023.
31. Taylor, R. et al. *GALACTICA: A Large Language Model for Science.* 2022.
32. Brown, T. B. et al. *Language Models Are Few-Shot Learners.* in *Proceedings of the 34th International Conference on Neural Information Processing Systems* (2020).

Motivation
ooooo

Background
oooo

Outline
oo

Corpus
oooooooooooo

Citations & Non-English
ooo

Artifact Parameters
oooooooooooooooooooo

Conclusion
oooooo

References

References V

33. Meyer, M., Frey, J., Laub, T., Wrzalik, M. & Krechel, D. *Citcom – Citation Recommendation*. in *INFORMATIK 2020* (eds Reussner, R. H., Koziolek, A. & Heinrich, R.) (Gesellschaft für Informatik, Bonn, 2021), 907–914. doi:10.18420/inf2020_82.
34. Parisot, M. & Zavrel, J. *Multi-objective Representation Learning for Scientific Document Retrieval*. in *Proceedings of the Third Workshop on Scholarly Document Processing* (Association for Computational Linguistics, Gyeongju, Republic of Korea, Oct. 2022), 80–88. <https://aclanthology.org/2022.sdp-1.9>.
35. Mochihashi, D. *Researcher2Vec: Neural Linear Model of Scholar Recommendation for Funding Agency*. in *Proceedings of the International Society of Scientometrics and Informetrics Conference 2023* (July 2023).
36. Veneri, M. D. et al. How Have Astronomers Cited Other Fields in the Last Decade? *Research Notes of the AAS* **6**, 113 (June 2022).
37. Xue, J. *An analysis of the semantic shifts of citations*. MA thesis (University of Helsinki, Faculty of Science, June 2021). <http://urn.fi/URN:NBN:fi:hulib-202107263435>.
38. Meng, X., Varol, O. & Barabási, A.-L. *Hidden Citations Obscure True Impact in Science*. in *Proceedings of the International Conference on Science of Science and Innovation 2023* (June 2023).
39. Raff, E. *A Step Toward Quantifying Independently Reproducible Machine Learning Research*. in *Advances in Neural Information Processing Systems* **32** (Curran Associates, Inc., 2019).
40. Sethi, A., Sankaran, A., Panwar, N., Khare, S. & Mani, S. *DLPaper2Code: Auto-Generation of Code From Deep Learning Research Papers*. *Proceedings of the AAAI Conference on Artificial Intelligence* **32**. doi:10.1609/aaai.v32i1.12326 (Apr. 2018).

Motivation
ooooo

Background
oooo

Outline
oo

Corpus
oooooooooooo

Citations & Non-English
ooo

Artifact Parameters
oooooooooooooooooooo

Conclusion
oooooo

References

References VI

41. Huang, K.-H., Tang, S. & Peng, N. *Document-level Entity-based Extraction as Template Generation*. in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (Nov. 2021), 5257–5269. doi:10.18653/v1/2021.emnlp-main.426.
42. Lee, S.-M. & Na, S.-H. *JBNU-CCLab at SemEval-2022 Task 12: Machine Reading Comprehension and Span Pair Classification for Linking Mathematical Symbols to Their Descriptions*. in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (July 2022), 1679–1686. doi:10.18653/v1/2022.semeval-1.231.
43. Popovic, N., Laurito, W. & Färber, M. *AIFB-WebScience at SemEval-2022 Task 12: Relation Extraction First - Using Relation Extraction to Identify Entities*. in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (July 2022), 1687–1694. doi:10.18653/v1/2022.semeval-1.232.
44. Baudart, G., Kirchner, P. D., Hirzel, M. & Kate, K. *Mining Documentation to Extract Hyperparameter Schemas*. in *Proceedings of the 7th ICML Workshop on Automated Machine Learning (AutoML 2020)* (2020).
45. Rak-Amnouykit, I., Milanova, A., Baudart, G., Hirzel, M. & Dolby, J. *Extracting Hyperparameter Constraints from Code*. in *ICLR Workshop on Security and Safety in Machine Learning Systems* (May 2021). <https://hal.science/hal-03401683> (2023).
46. Saier, T. & Färber, M. *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)* (Cologne, Germany, 2019), 14–26.
47. Saier, T. & Färber, M. *Semantic Modelling of Citation Contexts for Context-Aware Citation Recommendation*. in *Advances in Information Retrieval* (Springer International Publishing, 2020), 220–233. doi:10.1007/978-3-030-45439-5_15.

Motivation
ooooo

Background
oooo

Outline
oo

Corpus
oooooooooooo

Citations & Non-English
ooo

Artifact Parameters
oooooooooooooooooooo

Conclusion
oooooo

References

References VII

48. Krause, J., Shapiro, I., Saier, T. & Färber, M. *Bootstrapping Multilingual Metadata Extraction: A Showcase in Cyrillic*. in *Proceedings of the Second Workshop on Scholarly Document Processing* (Association for Computational Linguistics, Online, June 2021), 66–72. doi:10.18653/v1/2021.sdp-1.8. <https://aclanthology.org/2021.sdp-1.8>.
49. Shapiro, I., Saier, T. & Färber, M. *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).
50. Nishioka, C., Färber, M. & Saier, T. *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* (Association for Computing Machinery, 2022). ISBN: 9781450393454. doi:10.1145/3529372.3530953.
51. Saier, T., Dong, Y. & Färber, M. *CoCon: A Data Set on Combined Contextualized Research Artifact Use*. in *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (IEEE Computer Society, Los Alamitos, CA, USA, June 2023), 47–50. doi:10.1109/JCDL57899.2023.00016.

Motivation
ooooo

Background
oooo

Outline
oo

Corpus
oooooooooooo

Citations & Non-English
ooo

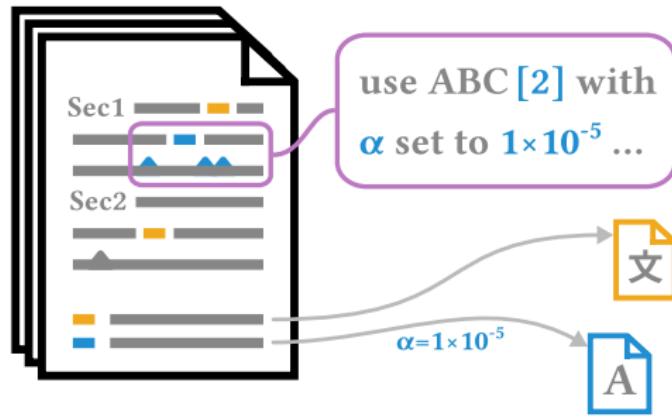
Artifact Parameters
oooooooooooooooooooo

Conclusion
oooooo

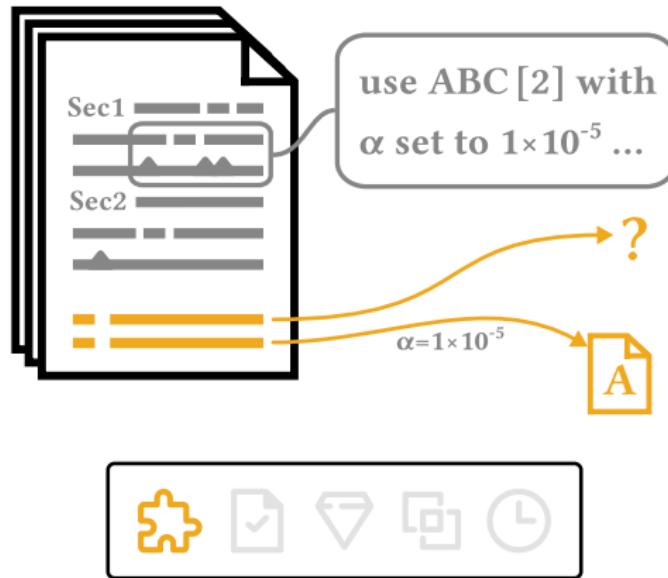
References

Extra Slides

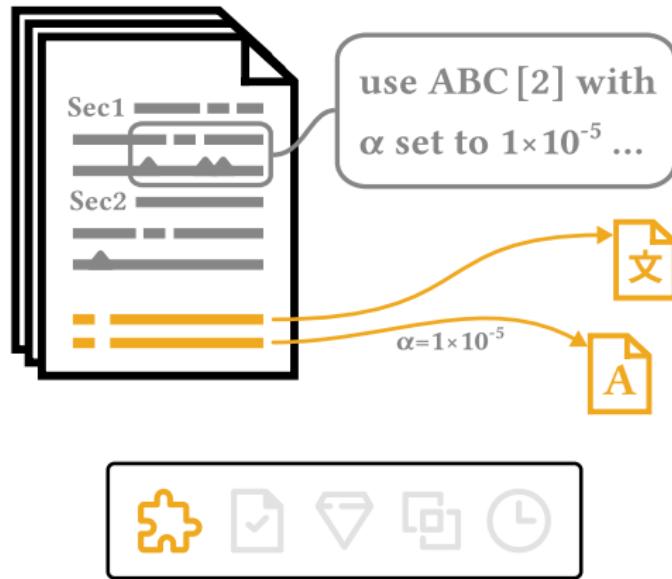
Quality Dimensions



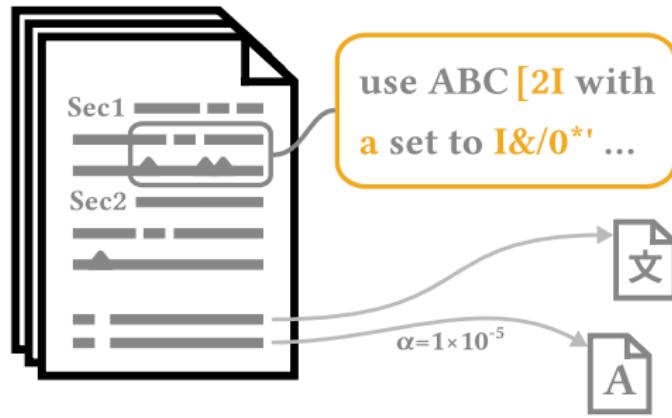
Quality Dimensions: Completeness



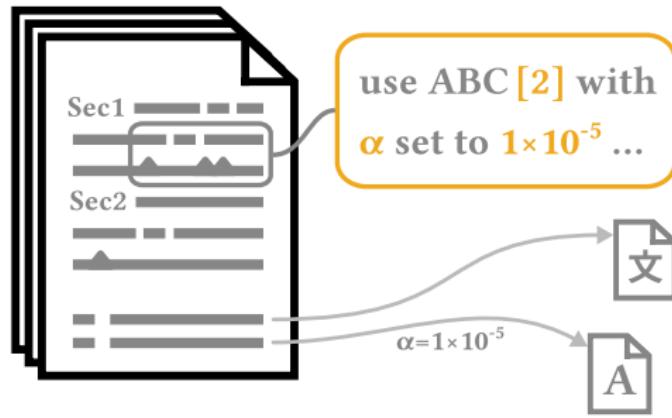
Quality Dimensions: Completeness



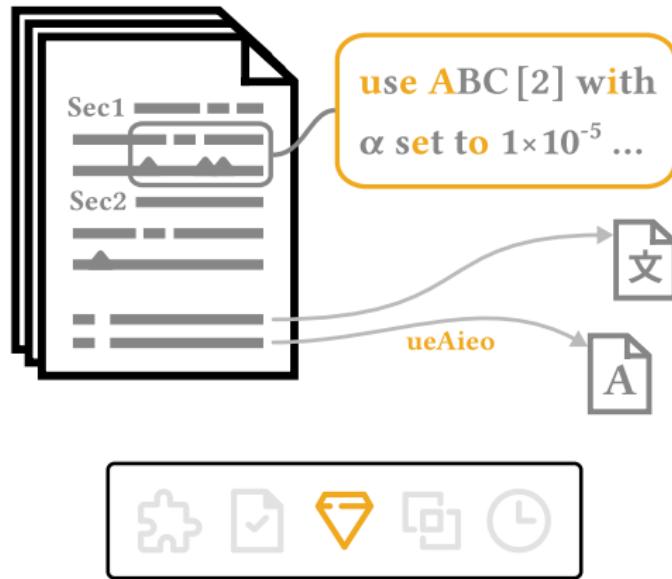
Quality Dimensions: Accuracy



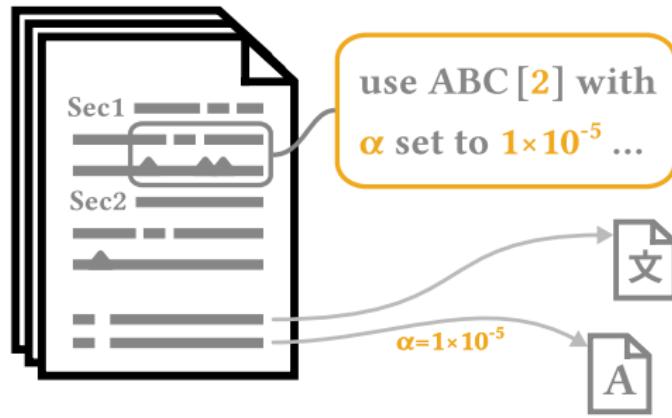
Quality Dimensions: Accuracy



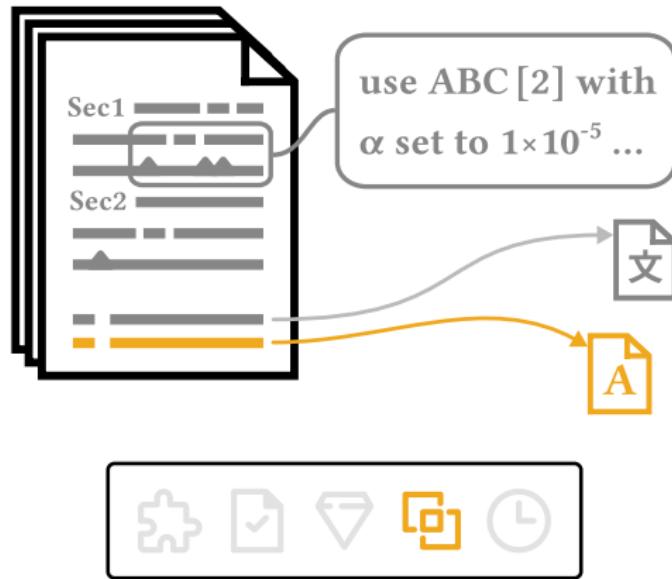
Quality Dimensions: Relevance



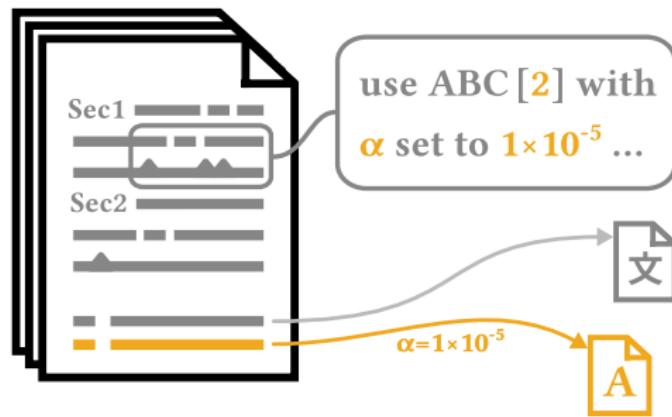
Quality Dimensions: Relevance



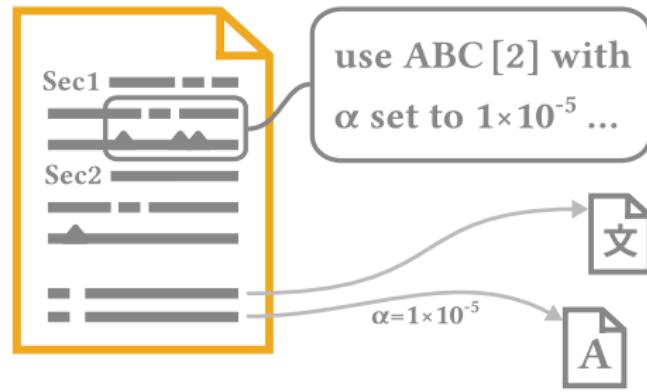
Quality Dimensions: Comparability



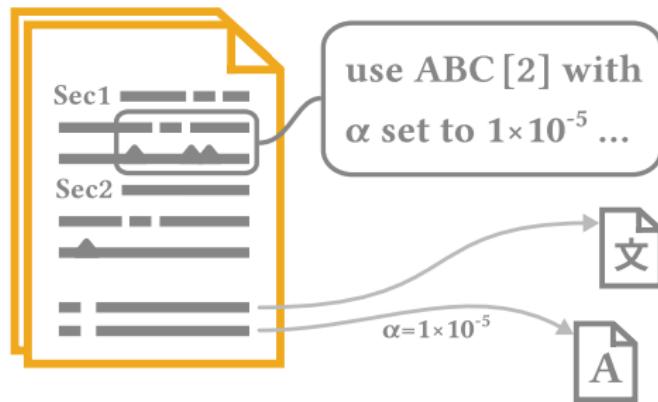
Quality Dimensions: Comparability



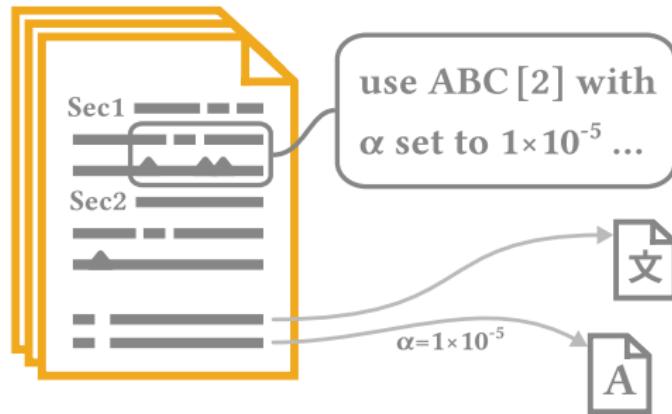
Quality Dimensions: Timeliness



Quality Dimensions: Timeliness



Quality Dimensions: Timeliness



Corpus - Result Full

Data set	# Docs	Cit. markers	Disciplines	Full text	Linked
ACL-ARC [2]	11 k	no	comp. ling.	PDF	✗
ACL-AAN [3]	18 k	no	comp. ling.	PDF	✗
Scholarly Dataset 2 [4]	100 k	no	CS	PDF	✗
CiteSeerX [5] / RefSeer [6]	1 M	ambiguous	mixed	400 char excerpts	✗
PMC OAS [7]	2.3 M	exact	biomedical	XML	mixed ^a
arXiv CS [8]	90 k	exact	CS	text	✓
unarXive [1]	1.2 M	exact	phys., maths, CS	text	✓

^a No citation network due to mixed set of IDs (PubMed, MEDLINE, DOI) [9].



Corpus - Result Full (2022)

Data Set	Source		Citation Network ^a			# Docs	Disciplines
	Data	Format	general	compare			
CORE [10]	multiple	PDF	0%	-	>100 M		various
S2ORC (PDF) [11]	multiple	PDF	69.4%	-	12 M		various
unarXive 2020 [1]	arXiv.org	L <small>A</small> T <small>E</small> X	42.6%	42.6%	1.2 M		phys., maths, CS
S2ORC (L <small>A</small> T <small>E</small> X) [11]	arXiv.org	L <small>A</small> T <small>E</small> X	31.1%	31.1%	1.5 M		phys., maths, CS
arXMLiv [12]	arXiv.org	L <small>A</small> T <small>E</small> X	0%	0%	1.6 M		phys., maths, CS
SciXGen [13]	arXiv.org	L <small>A</small> T <small>E</small> X	41.6%	-	205 k		CS
PMC-OAS [7]	PubMed	XML	mixed ^b	-	3.3 M		biomedical
unarXive 2022 [14]	arXiv.org	L <small>A</small> T <small>E</small> X	44.4%	44.4%	1.9 M		phys., maths, CS

^a “general”: all data; incomparable. “compare”: arXiv.org data 1991–2020; directly comparable.

^b No citation network due to mixed set of IDs (PubMed, MEDLINE, DOI) [9].



Corpus - Link Correctness Full

Table: Link correctness (n = 300)

Confidence level	Method ^a	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

^a Confidence interval given as Wilson score interval and Jeffreys interval [15].



Corpus - Stats

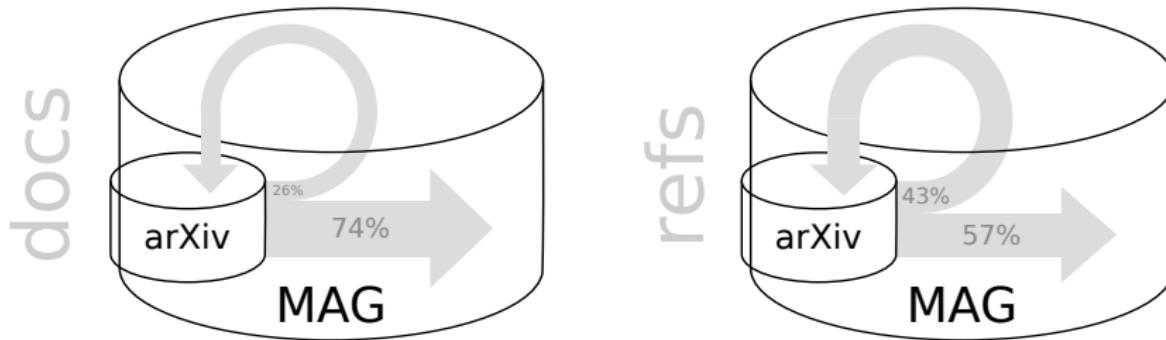
	citing documents	references	cited documents	
			outgoing	incoming
<i>full data set:</i>	1,043,126	15,954,664	15,954,664	2,746,288
full text	1,043,126	15,954,664	7,181,576	736,597
linked to MAG	994,351	15,846,351	15,954,664	2,746,288
<i>by discipline:</i>				
physics	662,894	9,300,576	7,827,072	921,852
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>

code: <https://github.com/IllDepence/unarXive>



Corpus - Citation Flow



Corpus - Reference Composition

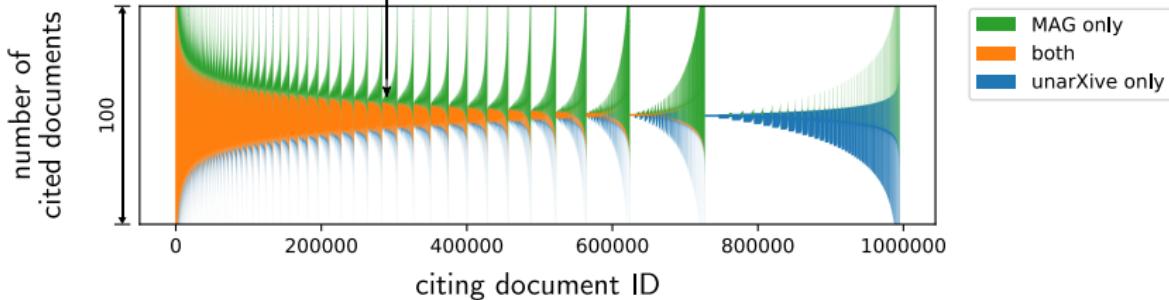
Exemplary visualization
of a single data point
on the x-axis.

document #288983

[...]

References

- [1] Example, A. (1970). Giving an example, *Journal of Examples*, 3, 11-27
- [2] Füller, F. (2019). Schnibberñe, *Zeitschrift der Schreibgeräte*, 77, 115-137
- [3] Beispiel, B. (1990). Beispielqualifikation, *Zeitschrift der Beispiele*, 1, 62-66
- [4] Tateo, T. (2015). Rönnji no Re, *Journal of Transcribed Titles*, 8, 8-30
- [5] Illustratio, I. (2017). Another Example, *Journal of Examples*, 40, 7-19
- [6] Hahn, H. (2018). The Last Word on Transcribed Titles, 18, 1-17
- [7] Pfeiffer, P. (2015). A Fair Article, *Transactions on Examples*, 18, 88-92
- [8] Sample, S. (2017). Yet Another Example, *Journal of Samples*, 79, 9-17
- [9] Example, E. (2011). Exemplaristik, *Zeitschrift der Exempli*, 13, 92-103
- [10] Owari, O. (2012). Sago Da, *Journal of Transcribed Titles*, 5, 357-381



Corpus - Target Sec. Specific Refs

	Discipline ^a	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	<u>Math[†]→Math[‡]</u>	200,859	5.41	6.81
	<u>Math[†]→CS</u>	5,134	92.13	2.96
	<u>Math[†]→Phys</u>	3,114	89.88	1.75
	<u>CS→Math[‡]</u>	3,456	18.82	0.41
	<u>Phys→Math[‡]</u>	3,859	16.49	0.40
	<u>CS→CS</u>	2,500	11.38	0.18
	<u>Phys→Phys</u>	10,374	2.12	0.14
	<u>CS→Phys</u>	50	307.16	0.10
	<u>Phys→CS</u>	137	101.40	0.09

^a \dagger : Mathematics citing document, \ddagger : Mathematics cited document, X→X: Citing and cited document are from the same discipline.



Artifact Parameters - Task

- Task Type
 - (Named) Entity Recognition, Relation Extraction

- 4 entity classes
 - (1) research **artifact**: model, method, data set, ...
 - descriptions of how authors use the artifacts
 - (2) **parameter** (α , learning rate, k, ...)
 - (3) **value** (1e-3, five, $\frac{1}{3}$, ...)
 - (4) **context** (for fine-tuning, during grid search, ...)

- 1 relation type
 - Given by entity type pair



Artifact Parameters - Task (ext)

- **Goal**

- Automatically extract hyperparameter information from paper text

- **Motivation**

- Reproducibility indication [39], automated reproduction [40]
 - Uncover conventions and trends
 - More fine-grained paper representations (similarity measures, recommendation, search)

- **Task Type**

- (Named) Entity Recognition, Relation Extraction



Artifact Parameters - Related Work: Fine-tuned models

- SciERC dataset
 - SCIEE [21]
 - PL-Marker [22]
 - → **entity type overlap**
- SciREX dataset
 - TempGen [41] (only RE)
- SemEval 2022 (math symb. to descr.)
 - JBUU-CCLab [42]
 - AIFB [43]
- SemEval 2021 (measurements)
 - LIORI [23]
 - → **utilize mention pattern regularities**

We evaluate our model on the task of **question answering** using

Section : Dataset

SQuAD is a **machine comprehension** dataset on a large set of **Wikipedia articles** , Two metrics are used to evaluate models: **Exact Match (EM)** and a softer metric , **F1 score**

Section: Model Details .

... Each paragraph and question are tokenized by a regular - expression - based word tokenizer (**PTB Tokenizer**) and fed into the model .
....

Section : Results .

The results of our model and competing approaches on the hidden test are summarized in Table [reference]. **BiDAF { ensemble }** achieves an **EM** score of 73.3 and an **F1**-score of 81.1 , outperforming all previous approaches .



Artifact Parameters - Related Work: Fine-tuned models (other data sources)

- From code documentation [44]
- From code [45]



Artifact Parameters - Approach: LLMs

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:



Artifact Parameters - Approach: LLMs

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?

Task

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

Input Text

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>  
/* [...] */  
...
```

Format

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.

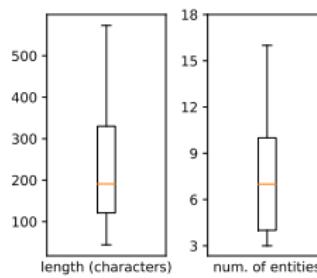
Output:

Output Prefix

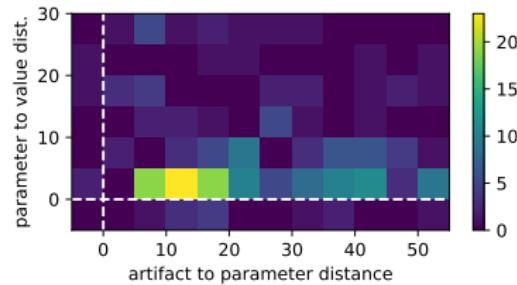


Artifact Parameters - Data

- Two annotation rounds
- Initial (pre-filtered text, exploratory, to fine-adjust scheme)
 - 151 text segments
 - 1,345 entities
 - 1,110 relations
- Main (full papers, eval data)
 - 444 paragraphs
 - 1,971 entities
(1,134 a, 131 p, 662 v, 44 c)
 - 614 relations
- IAA
 - 0.867 for entities
 - 0.737 for relations



(a) text segments



(b) relation distances (#chars)

Figure: Observations of initial annotation round

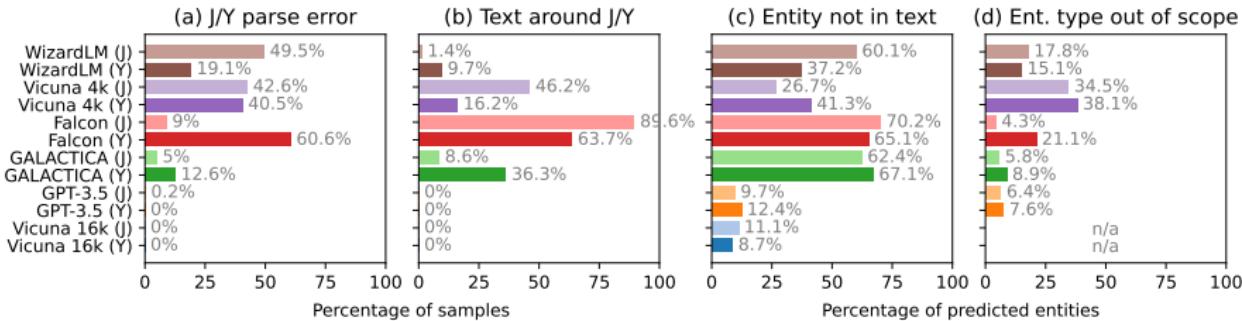


Artifact Parameters - Experiments: LLMs

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



Artifact Parameters - Experiments: LLMs



Publications - primary

Chap.	Venue	Year	Type	Length	Author Position	Venue Rating	Ref.
3	Scientometrics	2020	Journal	Full	1 of 2	SJR Q1	[1]
4	JCDL	2022	Workshop	Full	1 of 3	Core A*	[16]
	JCDL	2023	Conference	Short	1 of 3	Core A*	[14]
5	ICADL	2020	Conference	Full	1 of 2	Core A	[17]
	IJDL	2022	Journal	Full	1 of 3	SJR Q2	[18]
6	ECIR	2024	Conference	Full	1 of 4	Core A	[20]

Venue ranks from Core¹ (conferences) and SJR² (journals).³

¹See <http://portal.core.edu.au/conf-ranks/> (last accessed 2023-10-12).

²See <https://www.scimagojr.com/> (last accessed 2023-10-12).

³Ratings for publication year or, if not listed, most up-to-date ranking. Workshops ranks are that of the hosting conference.



Publications - secondary

Venue	Year	Type	Length	Author Position	Venue	Ref.
ECIR	2019	Workshop	Full	1 of 2	Core A	[46]
ECIR	2020	Conference	Full	1 of 3	Core A	[47]
NAACL	2021	Workshop	Short	3 of 4	Core A	[48]
AAAI	2022	Workshop	Full	2 of 3	Core A*	[49]
JCDL	2022	Conference	Full	3 of 3	Core A*	[50]
JCDL	2023	Conference	Short	1 of 3	Core A*	[51]

Additional publications (co-)authored leading up to and during the research period which are not a direct part of the dissertation, but nevertheless informed the overall research trajectory. Especially [46] and [48], which constitute the results of the master's thesis preceding the doctoral research period, paved the way for the dissertation.



Limitations

- **Corpus**

- L^AT_EX required (no humanities)

- **Citation Network**

- Blocking method scalability

- **Non-English Documents**

- Single “direction”
 - Dependency on author notation

- **Artifact Parameters**

- IE from text, not tables, code, etc.
 - ML specific
 - English only

