

# **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

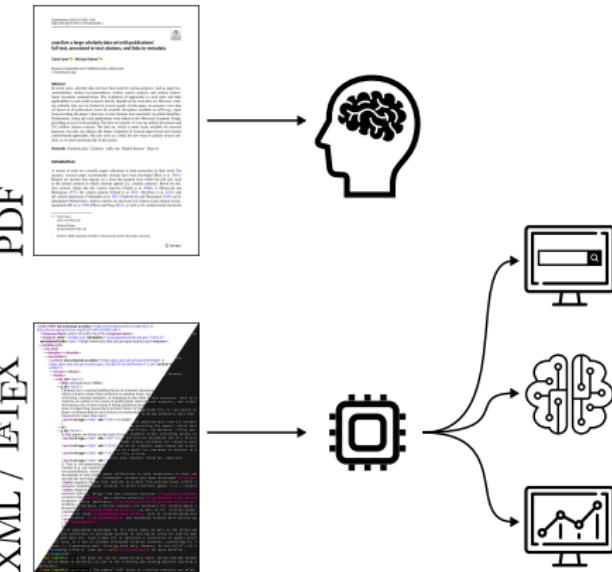
- **Services** (search, recommendation, statistics)
  - **ML models** (LLMs, summarization, recommender systems)
  - **Analyses** (temporal, geographic, institutional)

## ■ 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

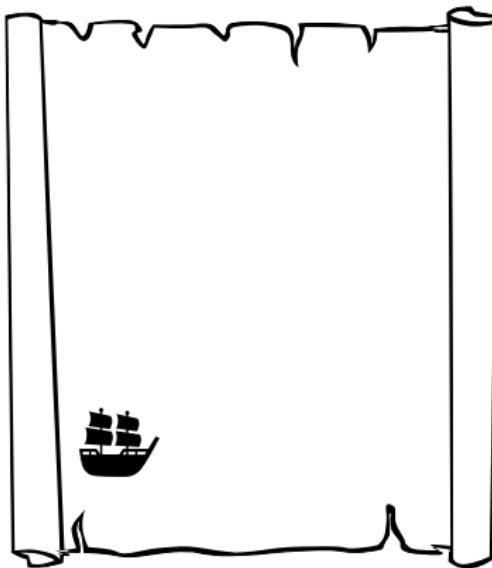
## Artifact Parameters

## Conclusion

## References

# Analogy

# Maps of the Sea



Motivation  
○○○●○

Background  
○○○○

Outline  
○○

Corpus  
○○○○○○○○○○

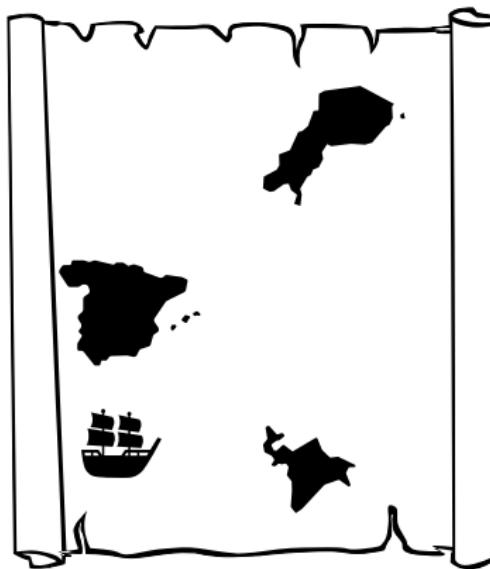
Citations & Non-English  
○○○

Artifact Parameters  
○○○○○○○○○○○○○○○○

Conclusion  
○○○○○○○○

References

# Maps of the Sea



## Motivation



## Background

## Outline

## Corpus



## Citations & Non-English ooo

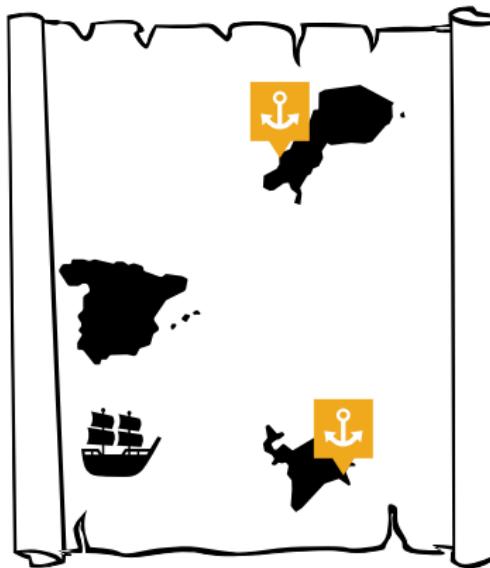
## Artifact Parameters

## Conclusion



## References

# Maps of the Sea



Motivation  
○○○●○

Background  
○○○○

Outline  
○○

Corpus  
○○○○○○○○○○

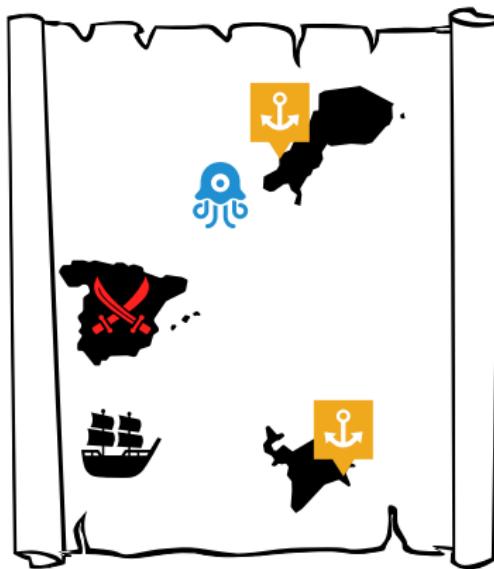
Citations & Non-English  
○○○

Artifact Parameters  
○○○○○○○○○○○○○○○○

Conclusion  
○○○○○○○○

References

# Maps of the Sea



## Motivation

○○○●○

## Background

## Outline

## Corpus



## Citations & Non-English ooo

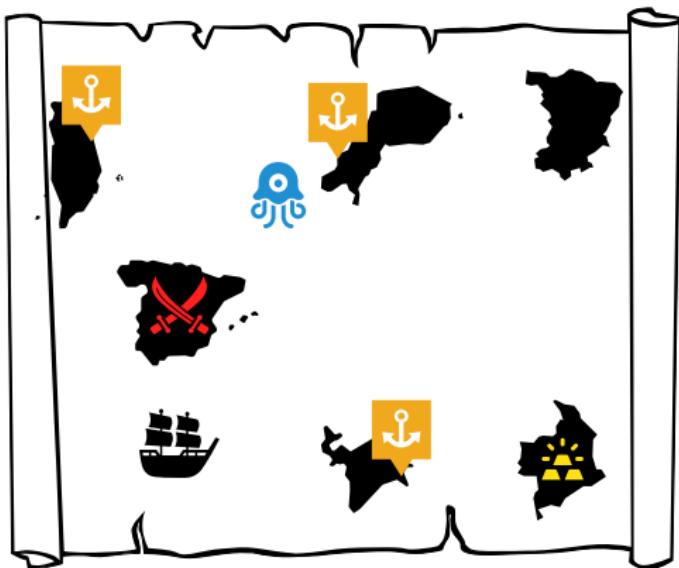
## Artifact Parameters

ooooooooooooooo

## Conclusion

## References

# Maps of the Sea



## Motivation



## Background

## Outline

## Corpus



## Citations & Non-English ooo

## Artifact Parameters

## Conclusion

oooooooo

## 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  
ooooooo

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  
ooo

Outline  
oo

Corpus  
oooooooo

Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooo

Conclusion  
ooooooo

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  
ooooooo

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.**

Motivation  
oooo●

Background  
ooo

Outline  
oo

Corpus  
oooooooo

Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooo

Conclusion  
ooooooo

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

**our maps of science are insufficient**

Motivation  
oooo●

Background  
oooo

Outline  
oo

Corpus  
oooooooo

Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooo

Conclusion  
ooooooo

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  
ooo

Outline  
oo

Corpus  
oooooooo

Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooo

Conclusion  
ooooooo

References

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

# Maps of Science



Motivation  
ooooo

Background  
○●○○

Outline  
○○

Corpus  
oooooooooo

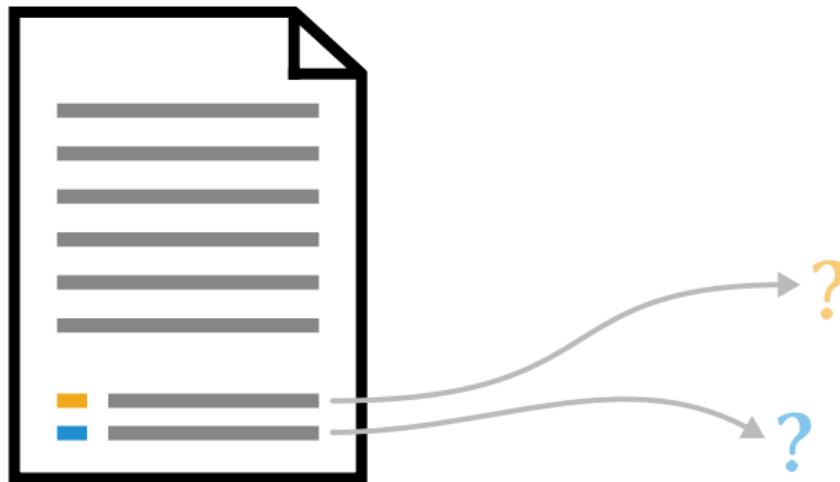
Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooo

Conclusion  
ooooooo

References

# Maps of Science



Motivation  
ooooo

Background  
●●○○

Outline  
○○

Corpus  
oooooooooo

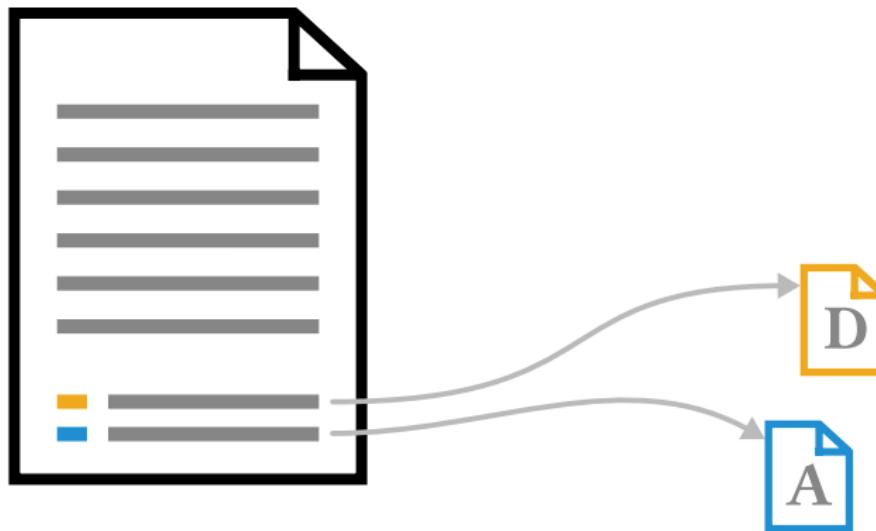
Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooooooo

Conclusion  
ooooooo

References

# Maps of Science



Motivation  
ooooo

Background  
●●○○

Outline  
○○

Corpus  
oooooooooo

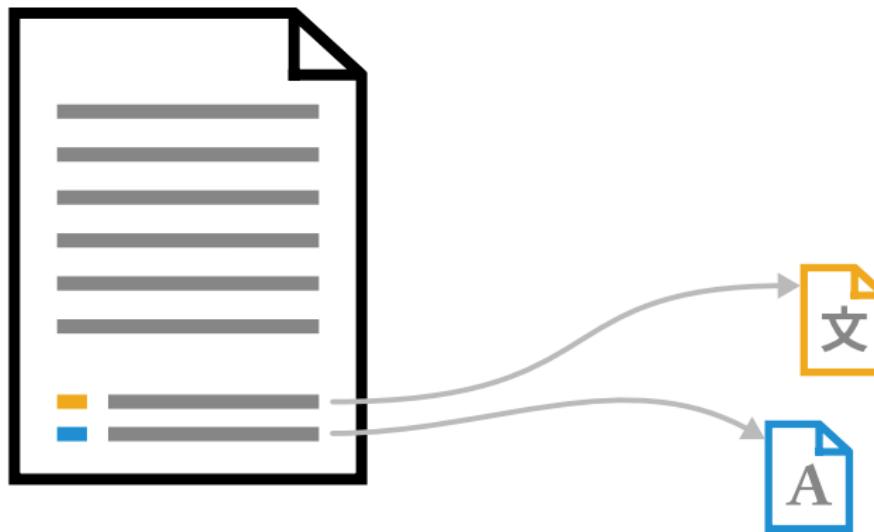
Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooooooo

Conclusion  
ooooooo

References

# Maps of Science



Motivation  
ooooo

Background  
●●○○

Outline  
○○

Corpus  
oooooooooo

Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooooooo

Conclusion  
ooooooo

References

# Maps of Science



Motivation  
ooooo

Background  
●●○○

Outline  
○○

Corpus  
oooooooooo

Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooooooo

Conclusion  
ooooooo

References

# Maps of Science



Motivation  
ooooo

Background  
●●○○

Outline  
○○

Corpus  
oooooooooo

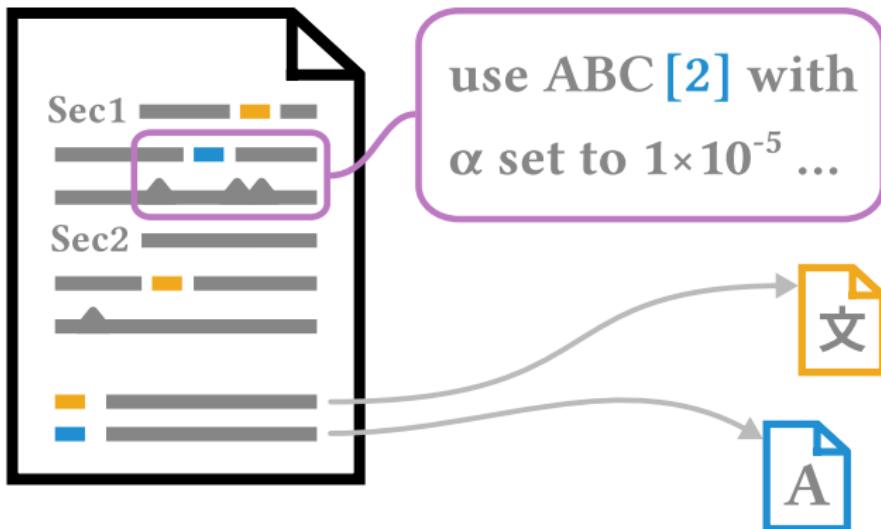
Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooooooo

Conclusion  
ooooooo

References

# Maps of Science



Motivation  
ooooo

Background  
●●○○

Outline  
○○

Corpus  
oooooooooo

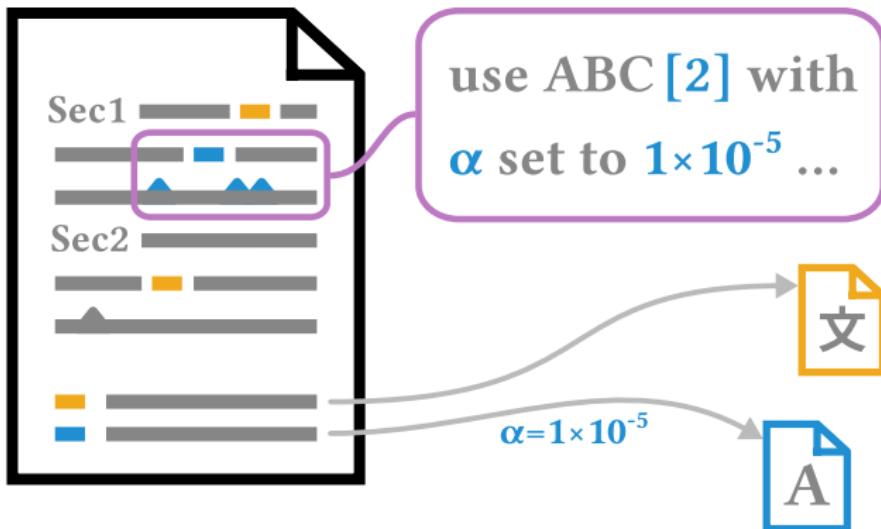
Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooooooo

Conclusion  
ooooooo

References

# Maps of Science



Motivation  
ooooo

Background  
●●○○

Outline  
○○

Corpus  
oooooooooo

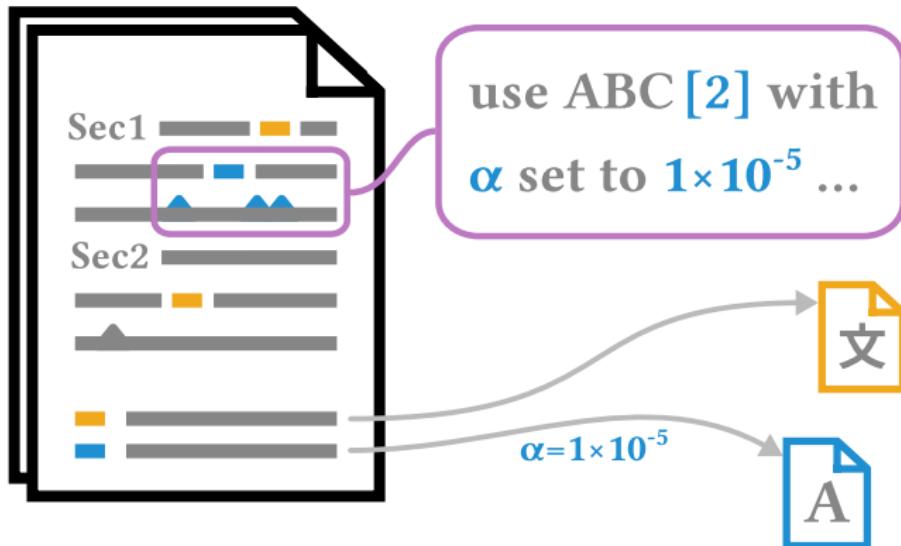
Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooooooo

Conclusion  
ooooooo

References

# Maps of Science



Motivation  
ooooo

Background  
●●○○

Outline  
○○

Corpus  
oooooooooo

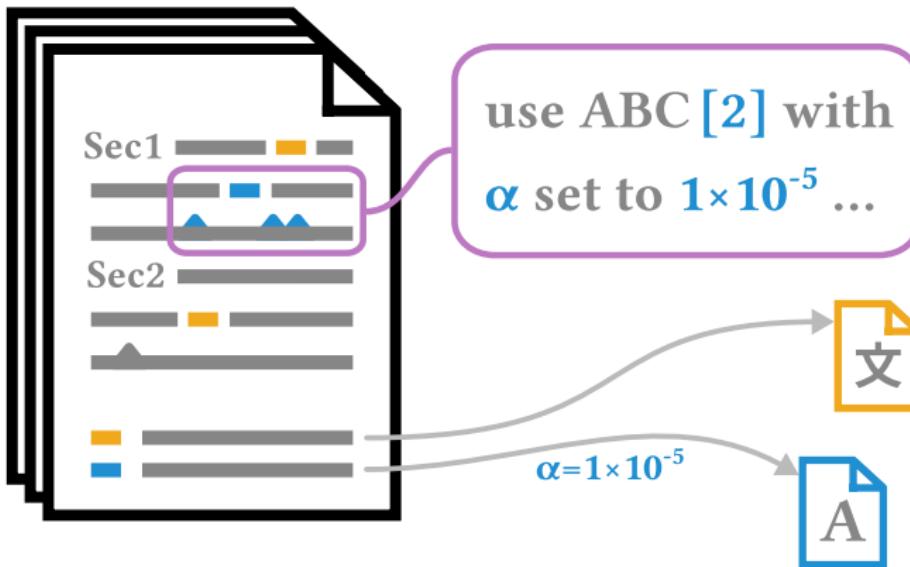
Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooooooo

Conclusion  
ooooooo

References

# Maps of Science



Motivation  
ooooo

Background  
●●○○

Outline  
○○

Corpus  
oooooooooo

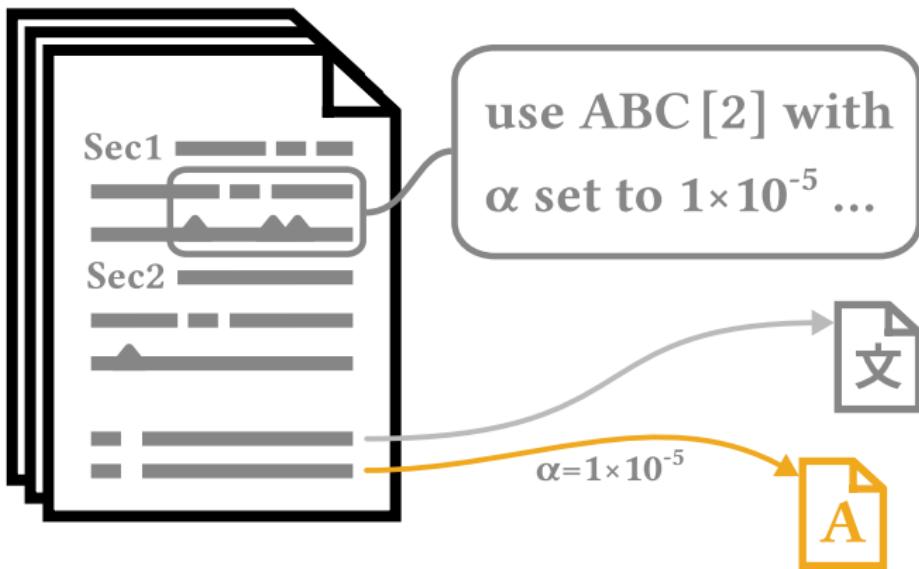
Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooooooo

Conclusion  
ooooooo

References

# Research Gap: Citation Network



Motivation  
ooooo

Background  
oo●o

Outline  
oo

Corpus  
oooooooooooo

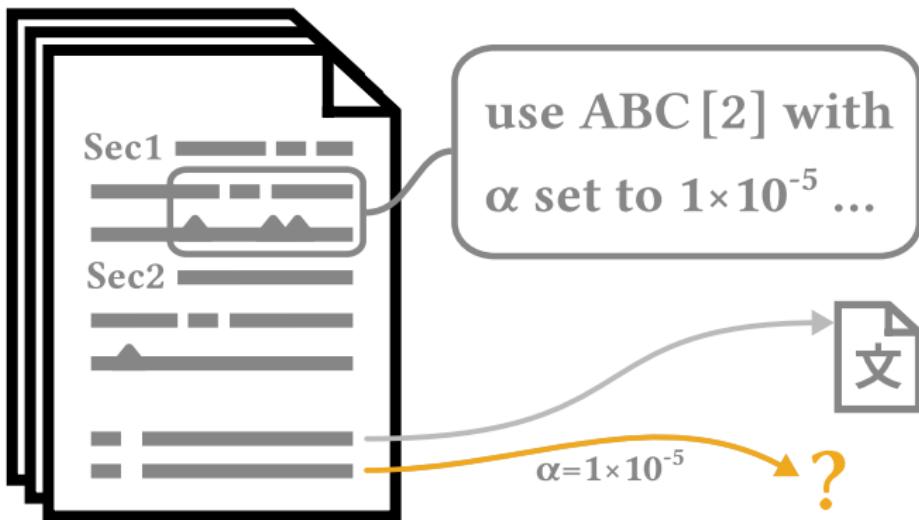
Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooooooooooo

Conclusion  
ooooooo

References

# Research Gap: Citation Network



Motivation  
ooooo

Background  
oo●○

Outline  
oo

Corpus  
oooooooooooo

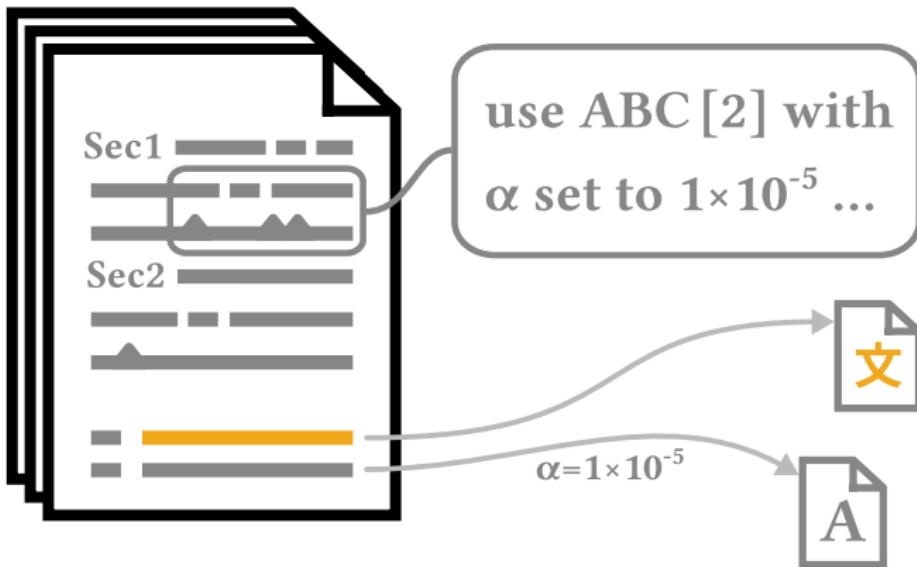
Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooooooooooo

Conclusion  
ooooooo

References

# Research Gap: Non-English Documents



Motivation  
ooooo

Background  
ooo●o

Outline  
oo

Corpus  
oooooooooooo

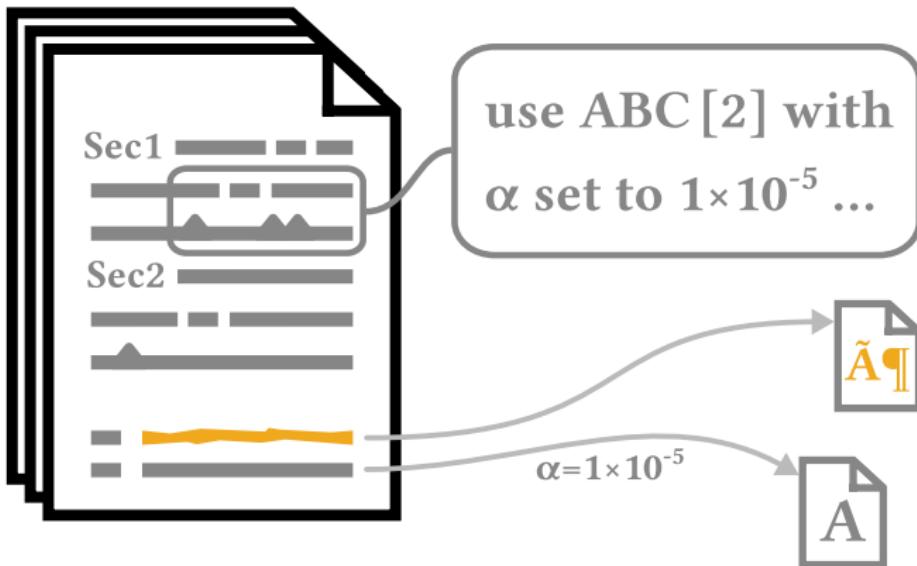
Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooooooooooo

Conclusion  
ooooooo

References

# Research Gap: Non-English Documents



Motivation  
ooooo

Background  
ooo●o

Outline  
oo

Corpus  
oooooooooooo

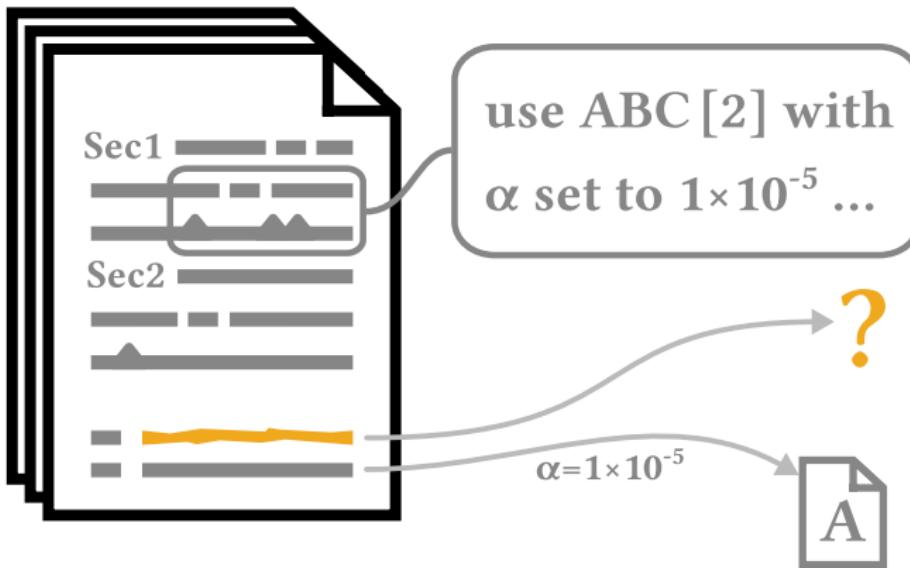
Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooooooooooo

Conclusion  
ooooooo

References

# Research Gap: Non-English Documents



Motivation  
ooooo

Background  
oo●○

Outline  
oo

Corpus  
oooooooooooo

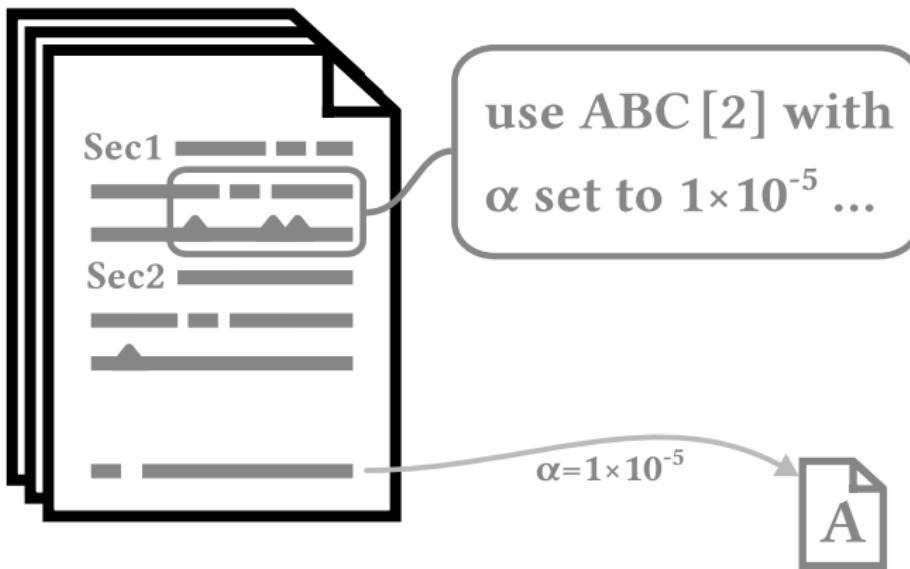
Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooooooooooo

Conclusion  
ooooooo

References

# Research Gap: Non-English Documents



Motivation  
ooooo

Background  
ooo●o

Outline  
oo

Corpus  
oooooooooooo

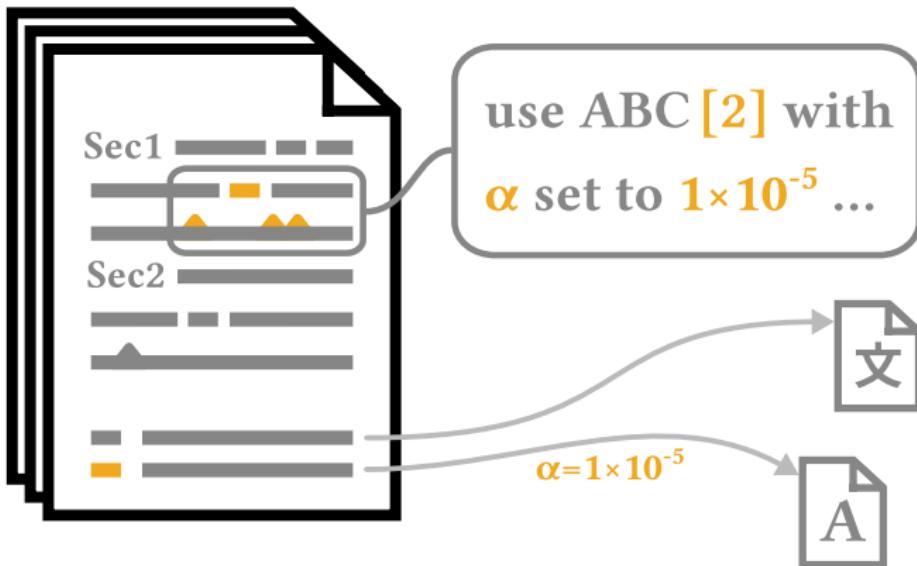
Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooooooooooo

Conclusion  
ooooooo

References

# Research Gap: Artifact Parameters



Motivation  
ooooo

Background  
ooo●o

Outline  
oo

Corpus  
oooooooooooo

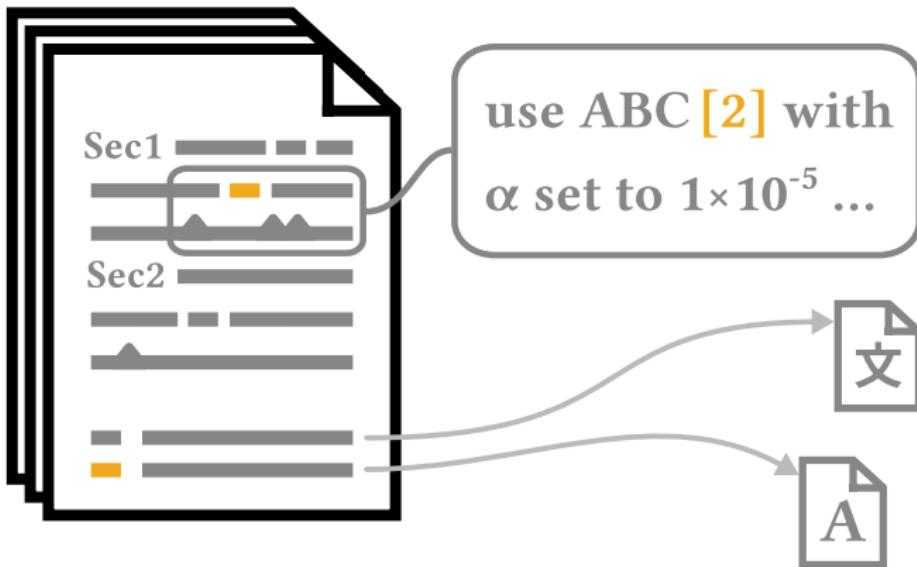
Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooooooooooo

Conclusion  
ooooooo

References

# Research Gap: Artifact Parameters



Motivation  
ooooo

Background  
oo●o

Outline  
oo

Corpus  
oooooooooooo

Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooooooooooo

Conclusion  
ooooooo

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



Base

Research Gap

RT1: Corpus

RT2: Citations

RT3: Non-English

RT4: Artifacts

Data Quality





## 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



## 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]

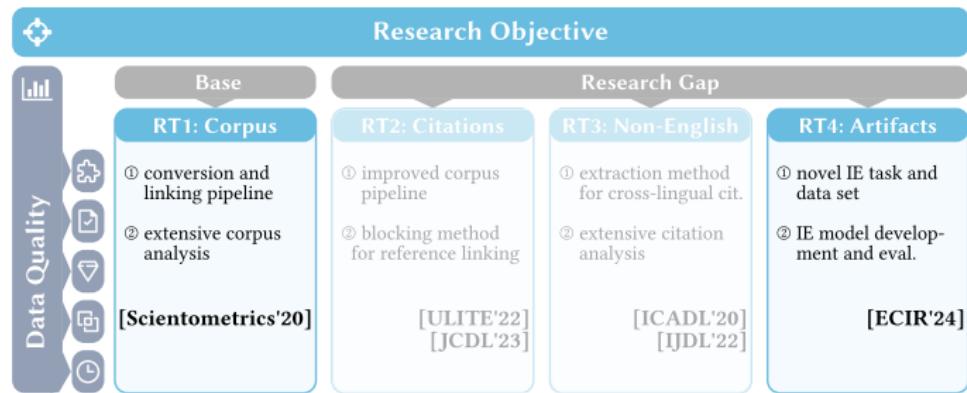
#### RT4: Artifacts

- 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  
oooooooooooo

Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooooooooooo

Conclusion  
ooooooo

References

# **Corpus**

**unarXive**

# Corpus - Digest

## ■ Research Task

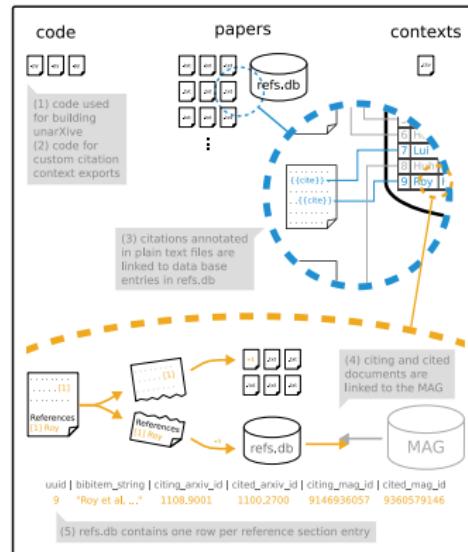
Develop a method for creating large-scale, high quality “base corpus”

## ■ Method

- L<sup>A</sup>T<sub>E</sub>X 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  
o●oooooooo

Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooo

Conclusion  
ooooooo

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
- $\text{\LaTeX}$  is powerful and people are creative

## ■ Reference linking

- Choice of target set
- Parsing (bb1, not bib)
- Variance and information sparsity

```
% Bonnet et al., 2001
\begin{document}
\newcommand{\nc}{\newcommand}
\nc{\be}{\begin{equation}}
```

[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  
ooo

Outline  
oo

Corpus  
oo●oooooooo

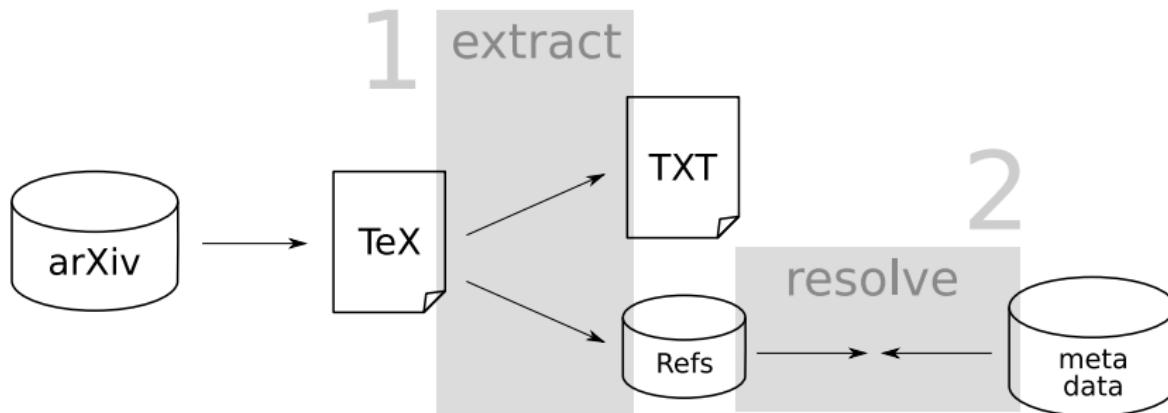
Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooo

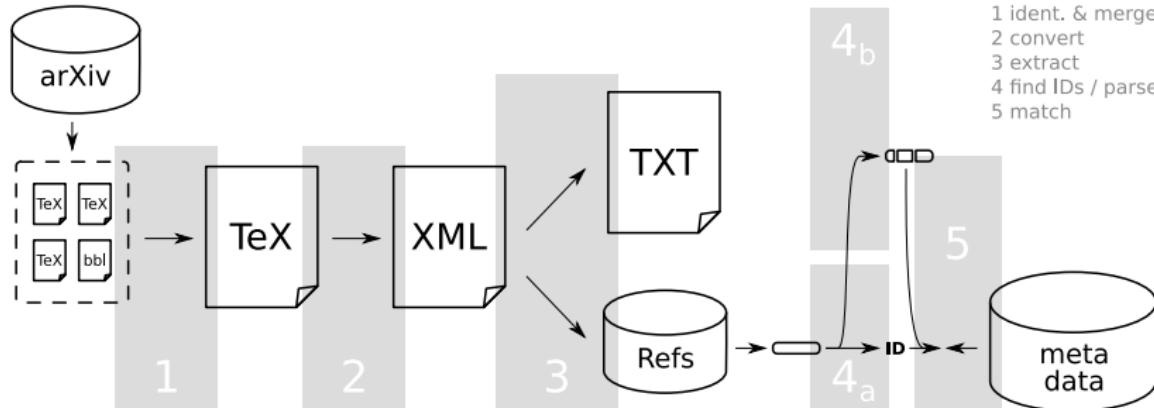
Conclusion  
ooooooo

References

# Corpus - Methods



# Corpus - Methods



# 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  
ooooooo

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]	<b>2.3 M</b>	exact	mixed <sup>a</sup>	-	-
arXiv CS [8]	90 k	exact	✓	39.3%	-
<b>unarXive [1]</b>	1.2 M	exact	✓	<b>42.6%</b>	<b>96%</b>

<sup>a</sup> 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]	<b>3.3 M</b>	exact	mixed <sup>a</sup>	-	-
SciXGen [13]	205 k	exact	✓	41.6%	-
S2ORC ( $\text{\LaTeX}$ ) [11]	1.5 M	exact	✓	31.1%	92%
<b>unarXive 2022 [14]</b>	1.9 M	exact	✓	<b>44.4%</b>	<b>96%</b>

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

Motivation  
ooooo

Background  
oooo

Outline  
oo

Corpus  
oooooooo●ooo

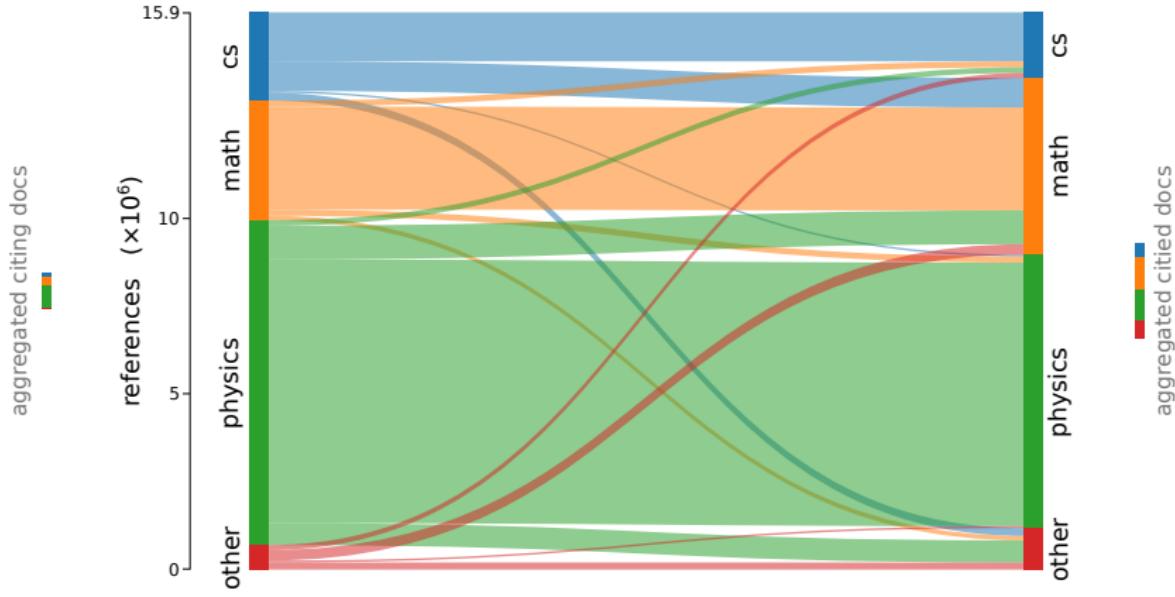
Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooo

Conclusion  
ooooooo

References

# Corpus - Citation Flow



Motivation  
ooooo

Background  
oooo

Outline  
oo

Corpus  
oooooooo●○

Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooo

Conclusion  
ooooooo

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<sub>LATEX</sub>)
  - 4% more accurate reference links ( $\leftrightarrow$  S2ORC<sub>LATEX</sub>)
- 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  
oooooooooooo

Conclusion  
ooooooo

References

# **Citations & Non-English**

# Citation Network - Digest

## ■ Research Task

Develop a method linking references more successfully without compromising accuracy

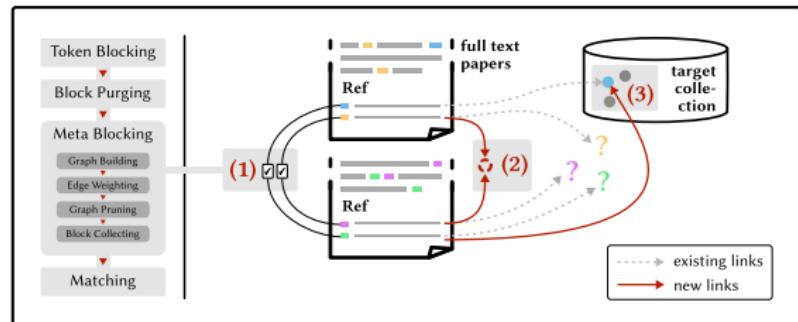
## ■ Method

- 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  
ooo

Outline  
oo

Corpus  
oooooooooooo

Citations & Non-English  
○●○

Artifact Parameters  
oooooooooooooooo

Conclusion  
ooooooo

References

# Non-English Documents - Digest

## ■ Research Task

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

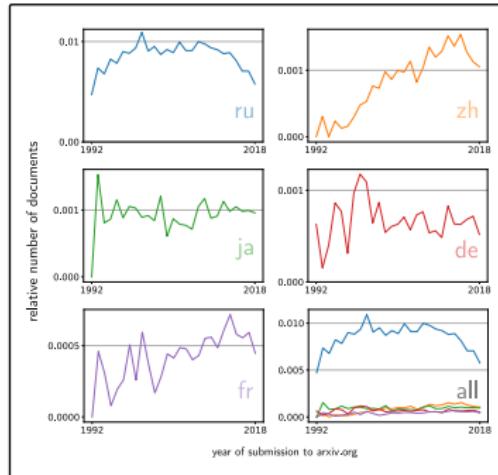
## ■ Method

- 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  
ooo

Outline  
oo

Corpus  
oooooooooooo

Citations & Non-English  
oo●

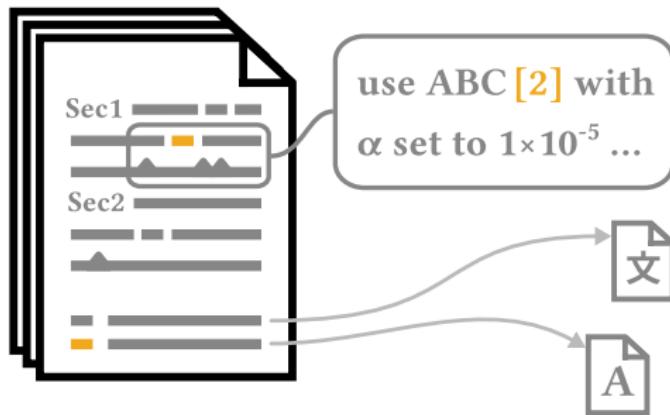
Artifact Parameters  
oooooooooooooooooooo

Conclusion  
ooooooo

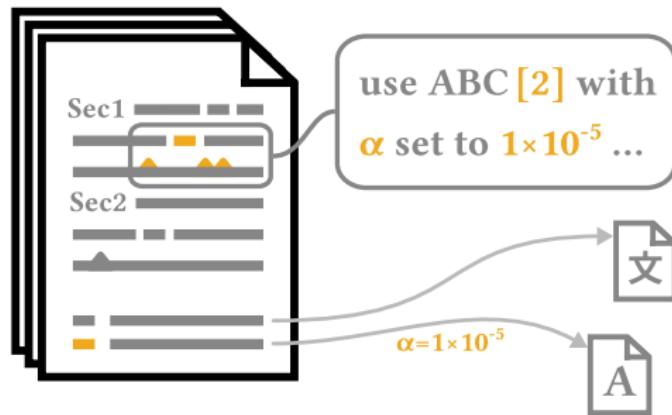
References

# **Artifact Parameters**

# Artifact Parameters

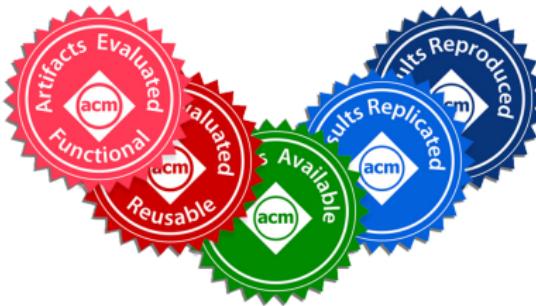


# Artifact Parameters



# Artifact Parameters - Motivation

The image shows two screenshots of scientific platforms. On the left, the D4Science homepage displays a "Trending Research" section featuring a paper titled "Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction". The paper has 1,386 citations and was published by FoundationVision/VRL on April 3, 2024. It includes links to the paper and code. On the right, the "Dataset Search" interface from Google shows a search bar with the placeholder "Search for Datasets" and a link to "Learn more about Dataset Search".



Motivation  
ooooo

Background  
oooo

Outline  
oo

Corpus  
oooooooo

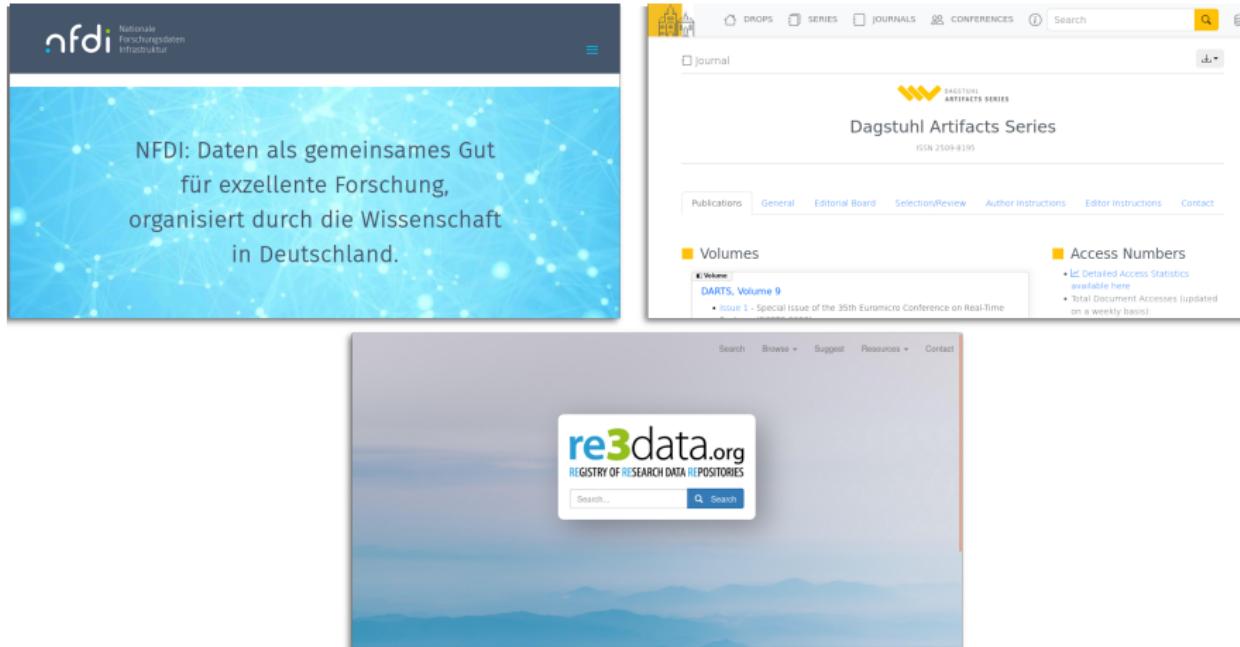
Citations & Non-English  
ooo

Artifact Parameters  
○○●oooooooooooo

Conclusion  
ooooooo

References

# Artifact Parameters - Motivation



Motivation  
ooooo

Background  
oooo

Outline  
oo

Corpus  
oooooooooooo

Citations & Non-English  
ooo

Artifact Parameters  
○○●oooooooooooo

Conclusion  
ooooooo

References

# Artifact Parameters - Motivation

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

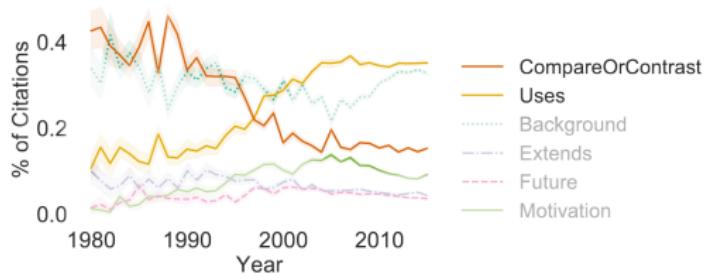


Figure 5: Changes in the average citation frame in ACL papers reveals a continued decline in the percentage of COMPARISON OR CONTRAST and increase in USES citations. The increase in BACKGROUND citations circa 2010 marks the start of the era of unlimited references in ACL conferences. Shaded regions show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

Motivation  
ooooo

Background  
oooo

Outline  
oo

Corpus  
oooooooo

Citations & Non-English  
ooo

Artifact Parameters  
○○●oooooooooooo

Conclusion  
ooooooo

References

# Artifact Parameters - Digest

## ■ Research Task

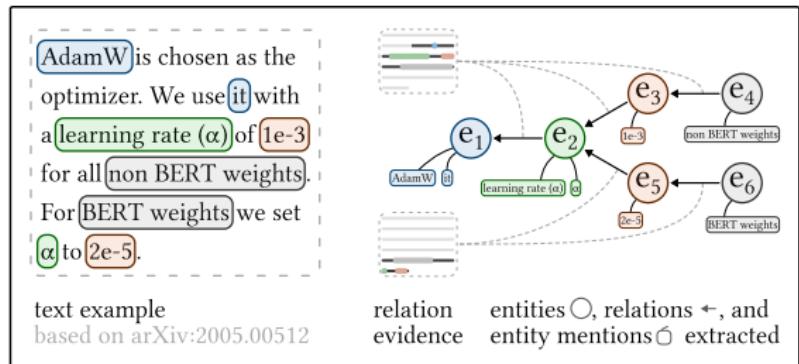
Develop a method to extract fine-grained research artifacts information

## ■ 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  
oooooooooooo

Citations & Non-English  
ooo

Artifact Parameters  
oooo●oooooooooooo

Conclusion  
ooooooo

References

# Artifact Parameters - Task

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

text example  
based on arXiv:2005.00512

Motivation  
ooooo

Background  
oooo

Outline  
oo

Corpus  
oooooooo

Citations & Non-English  
ooo

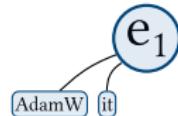
Artifact Parameters  
oooo●oooooooo

Conclusion  
ooooooo

References

# Artifact Parameters - Task

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

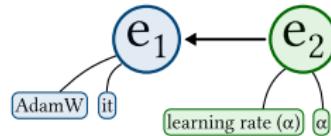


text example  
based on arXiv:2005.00512

entity type  
artifact

# Artifact Parameters - Task

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



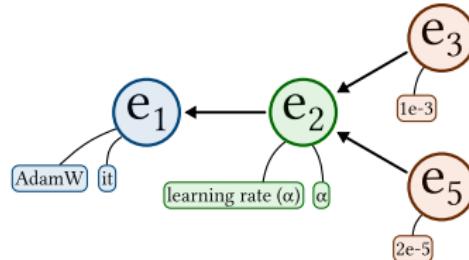
text example  
based on arXiv:2005.00512

entity type  
artifact parameter

# Artifact Parameters - Task

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

text example  
based on arXiv:2005.00512

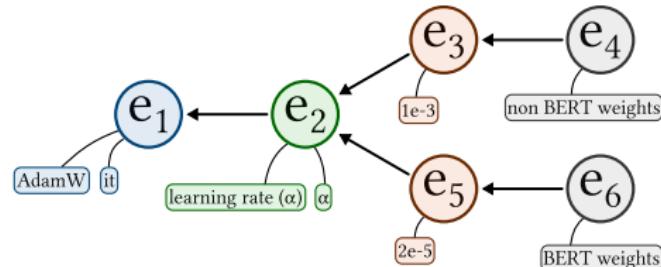


entity type  
artifact parameter value

# Artifact Parameters - Task

AdamW is chosen as the optimizer. We use it with a learning rate ( $\alpha$ ) of  $1e-3$  for all non BERT weights. For BERT weights we set  $\alpha$  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 [21]
- SOTA: PL-Marker [22]  
→ **entity type overlap**

## ■ LLMs

- Mostly singular values/lists [24–26]
- Hierarchical [27] (see right)  
→ **data serialization format**

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

- SciERC dataset [21]
- SOTA: PL-Marker [22]
- **entity type overlap**

## ■ LLMs

- Mostly singular values/lists [24–26]
- Hierarchical [27] (see right)
- **data serialization format**

Note: all related LLM work is evaluated on **GPT models only**.

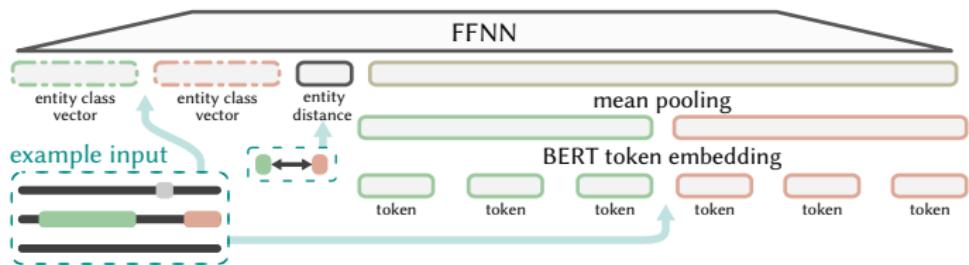
**Document:** The charge and discharge performance of an all-solid-state lithium battery with the LiBH4-LiI solid solution as an electrolyte is reported. Lithium titanate (Li<sub>4</sub>Ti<sub>5</sub>O<sub>12</sub>) was used as the positive electrode and...

### List of json documents

```
[ { "formula": "LiBH4-LiI", "description": "solid solution", "application": ["Li-ion battery", "electrolyte"] }, { "name": "lithium titanate", "formula": "Li4Ti5O12", "application": ["Li-ion battery", "positive electrode"] } ]
```

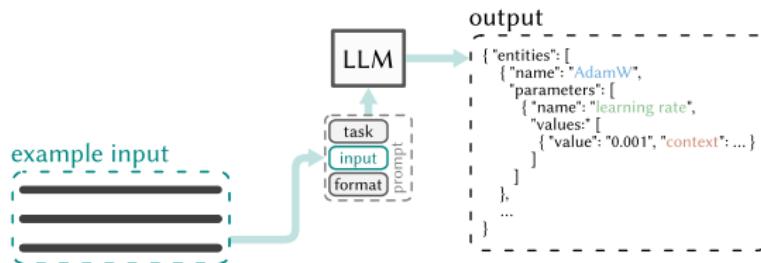
# Approach: Fine-tuned models

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



# Approach: LLMs

- Data serialization format
  - JSON → YAML
- Compare 6 models
- Base prompt + tuning for each
- Zero-shot: all
- Few-shot: only Vicuna<sub>16k</sub>



# Approach: LLMs

- Data serialization format
  - JSON → YAML
- Compare 6 models
- Base prompt + tuning for each
- Zero-shot: all
- Few-shot: only Vicuna<sub>16k</sub>

Model	Size
WizardLM [28]	13 B
Vicuna <sub>4k</sub> [29]	13 B
Vicuna <sub>16k</sub> [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●ooooo

Conclusion  
ooooooo

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

oooooo

## 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  
ooooooo

References

# Artifact Parameters - Experiments: Fine-tuned models

## ■ Setting

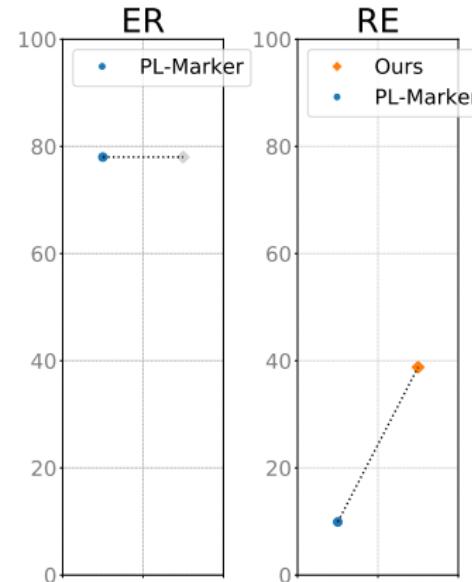
- 5-fold cross-validation
- Stratified sampling

## ■ 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●oooo

Conclusion  
ooooooo

References

# Artifact Parameters - Experiments: LLMs

## ■ Setting

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

## ■ Results (best F<sub>1</sub> [%]) (zero-shot)

- ER: 44.0 (37.4)
- RE: 7.8 (6.1)

## ■ Analysis

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

Motivation  
ooooo

Background  
oooo

Outline  
oo

Corpus  
oooooooo

Citations & Non-English  
ooo

Artifact Parameters  
oooooooo●oo

Conclusion  
ooooooo

References

# Artifact Parameters - Experiments: LLMs

## Setting

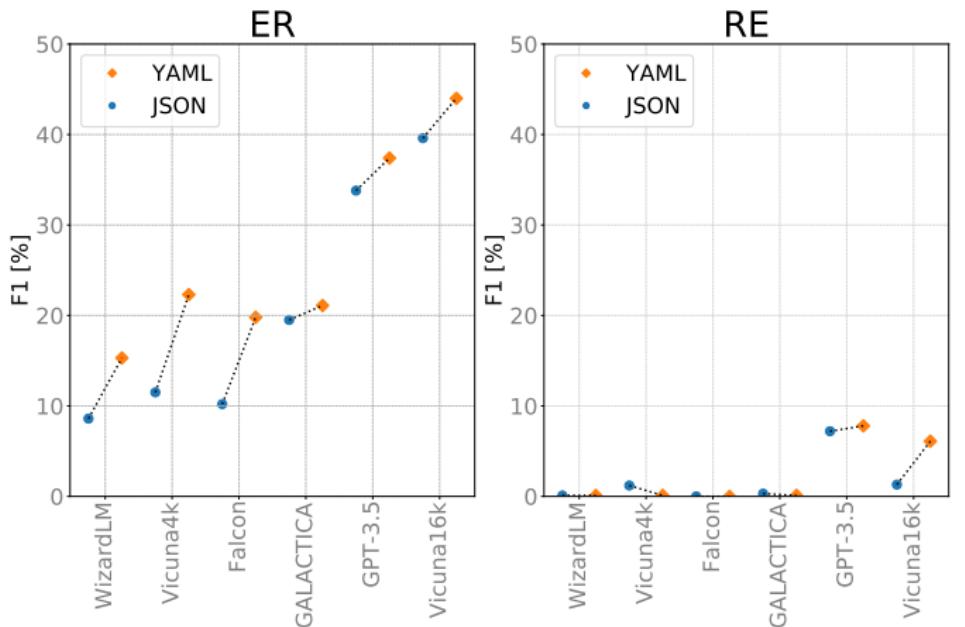
- Zero-/Few-shot
- Compare JSON w/ YAML variant

## Results (best F<sub>1</sub> [%]) (zero-shot)

- ER: 44.0 (37.4)
- RE: 7.8 (6.1)

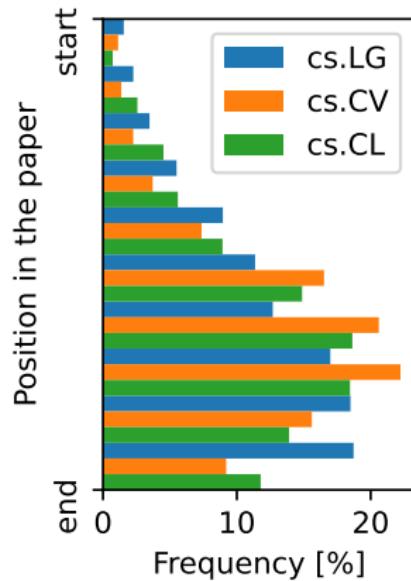
## Analysis

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



# 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  
oooooooooooo

Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooo●○

Conclusion  
ooooooo

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  
oooooooooooo

Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooo●

Conclusion  
ooooooo

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 ✓



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 referentie 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.

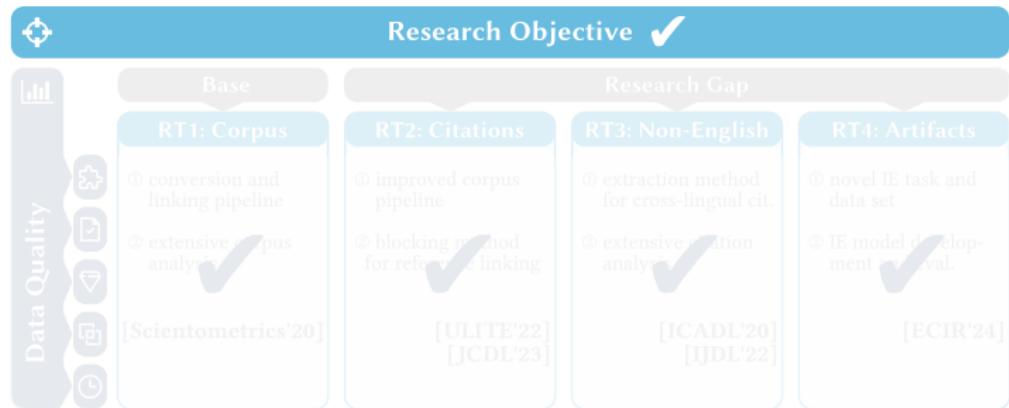
[ECIR'24]

# 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  
ooo

Outline  
oo

Corpus  
oooooooooooo

Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooooooooooo

Conclusion  
oo●oooo

References

# Impact

Adoption by the research community.

## ■ Methodology

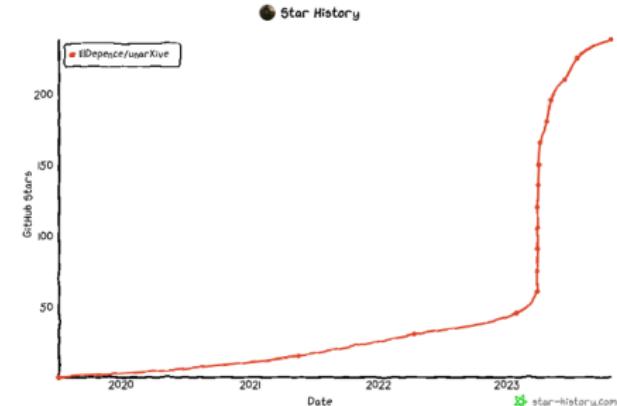
- Document Parsing Methodology [11]
- Reference Matching Methodology [13]

## ■ 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]



Motivation  
ooooo

Background  
oooo

Outline  
oo

Corpus  
oooooooooooo

Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooooooooooo

Conclusion  
oooo●oooo

References

# 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  
oooooooooooo

Citations & Non-English  
ooo

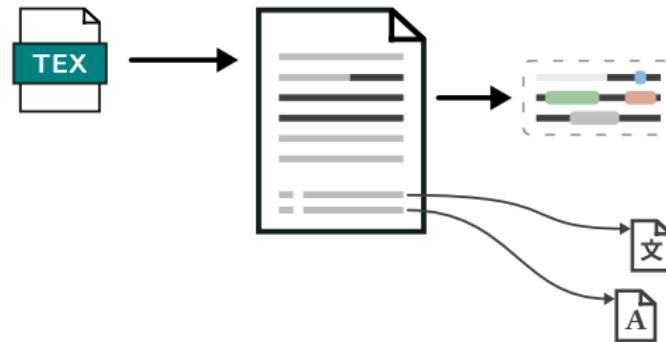
Artifact Parameters  
oooooooooooooooooooo

Conclusion  
oooo●ooo

References

# Outlook

- **Data Sources**
  - JATS-XML
- **Non-English Publications**
  - English citing docs
- **Artifact Parameters**
  - Applications
- **Long-term**
  - Digital record of science



Motivation  
ooooo

Background  
oooo

Outline  
oo

Corpus  
oooooooooooo

Citations & Non-English  
ooo

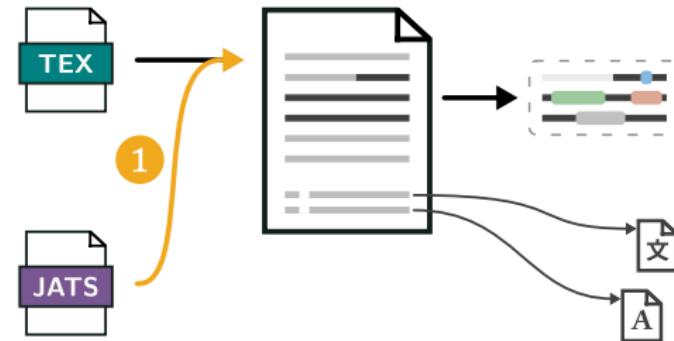
Artifact Parameters  
oooooooooooooooooooo

Conclusion  
oooooo●oo

References

# Outlook

- **Data Sources**
  - JATS-XML
- **Non-English Publications**
  - English citing docs
- **Artifact Parameters**
  - Applications
- **Long-term**
  - Digital record of science



Motivation  
ooooo

Background  
oooo

Outline  
oo

Corpus  
oooooooooooo

Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooooooooooo

Conclusion  
oooooo●oo

References

# Outlook

- **Data Sources**
  - JATS-XML
- **Non-English Publications**
  - English citing docs
- **Artifact Parameters**
  - Applications
- **Long-term**
  - Digital record of science



Motivation  
ooooo

Background  
ooo

Outline  
oo

Corpus  
oooooooooooo

Citations & Non-English  
ooo

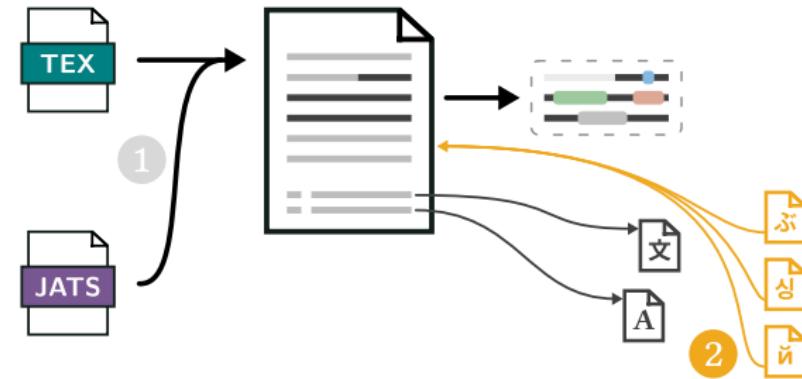
Artifact Parameters  
oooooooooooooooooooo

Conclusion  
oooooo●oo

References

# Outlook

- **Data Sources**
  - JATS-XML
- **Non-English Publications**
  - English citing docs
- **Artifact Parameters**
  - Applications
- **Long-term**
  - Digital record of science



Motivation  
ooooo

Background  
ooo

Outline  
oo

Corpus  
oooooooooooo

Citations & Non-English  
ooo

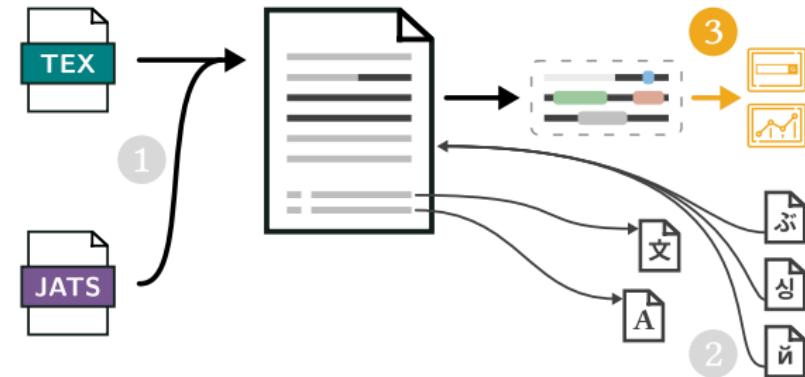
Artifact Parameters  
oooooooooooooooooooo

Conclusion  
oooooo●oo

References

# Outlook

- **Data Sources**
  - JATS-XML
- **Non-English Publications**
  - English citing docs
- **Artifact Parameters**
  - Applications
- **Long-term**
  - Digital record of science



Motivation  
ooooo

Background  
ooo

Outline  
oo

Corpus  
oooooooooooo

Citations & Non-English  
ooo

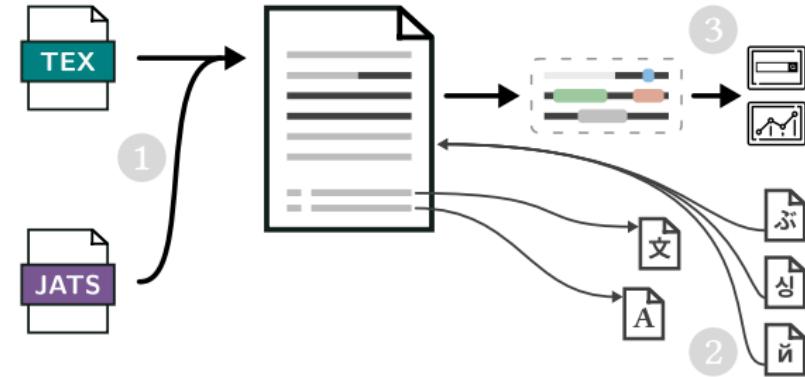
Artifact Parameters  
oooooooooooooooooooo

Conclusion  
oooooo●oo

References

# Outlook

- **Data Sources**
  - JATS-XML
- **Non-English Publications**
  - English citing docs
- **Artifact Parameters**
  - Applications
- **Long-term**
  - Digital record of science



Motivation  
ooooo

Background  
ooo

Outline  
oo

Corpus  
oooooooooooo

Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooooooooooo

Conclusion  
oooooo●oo

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  
ooooooo

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  
ooooooo

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  
ooo

Outline  
oo

Corpus  
oooooooooooo

Citations & Non-English  
ooo

Artifact Parameters  
oooooooooooooooooooo

Conclusion  
ooooooo

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  
ooooooo

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  
ooooooo

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  
ooooooo

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  
ooo

Outline  
oo

Corpus  
oooooooooooo

Citations & Non-English  
ooo

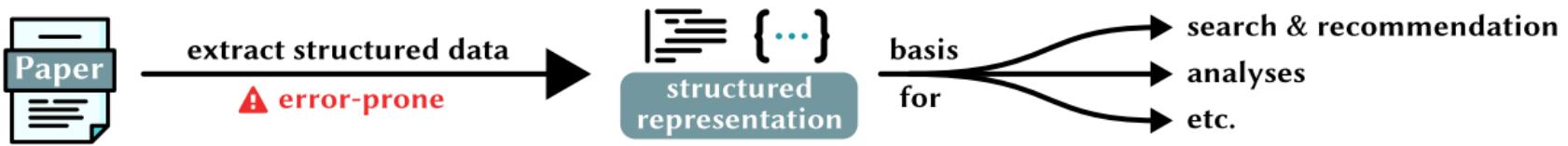
Artifact Parameters  
oooooooooooooooooooo

Conclusion  
ooooooo

References

# **Extra Slides**

# Scholarly Data IE



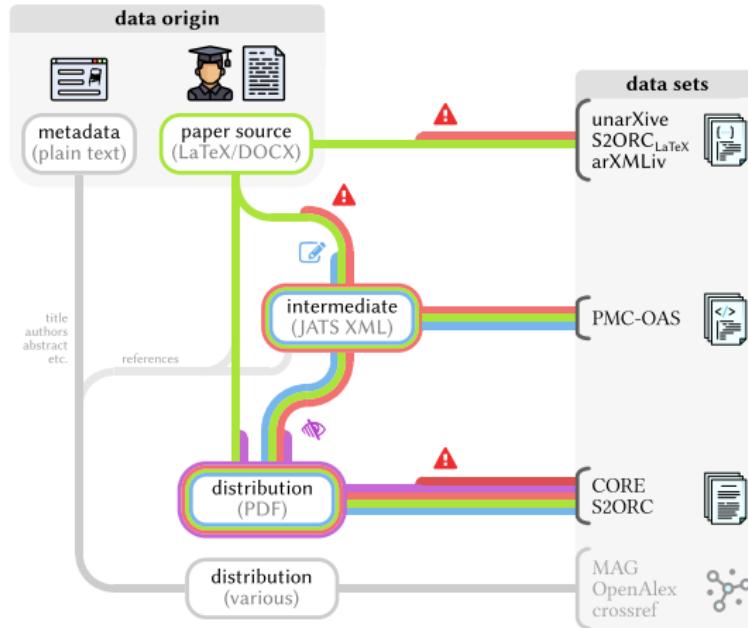
Scholarly Data  
●○

Corpus  
oooooooo

Artifact Parameters  
oooooooooooo

Misc  
oooo

# Scholarly Data Origins



# 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 <sup>a</sup>
arXiv CS [8]	90 k	exact	CS	text	✓
<b>unarXive</b> [1]	1.2 M	exact	phys., maths, CS	text	✓

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

# Corpus - Result Full (2022)

Data Set	Source		Citation Network <sup>a</sup>			# 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 <sup>b</sup>	-	<b>3.3 M</b>		<b>biomedical</b>
<b>unarXive 2022 [14]</b>	arXiv.org	L <small>A</small> T <small>E</small> X	<b>44.4%</b>	<b>44.4%</b>	<b>1.9 M</b>		<b>phys., maths, CS</b>

<sup>a</sup> “general”: all data; incomparable. “compare”: arXiv.org data 1991–2020; directly comparable.

<sup>b</sup> 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 <sup>a</sup>	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

<sup>a</sup> Confidence interval given as Wilson score interval and Jeffreys interval [15].

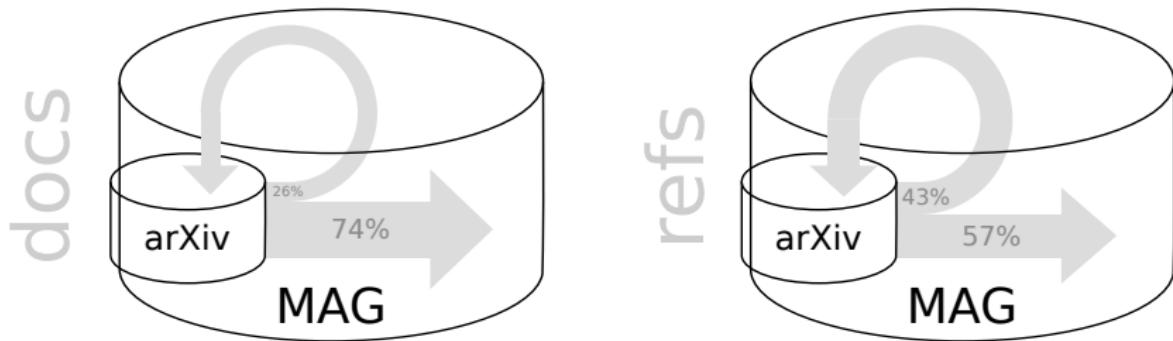
# Corpus - Stats

	citing documents	references	cited documents	
			outgoing	incoming
<i>full data set:</i>	<b>1,043,126</b>	<b>15,954,664</b>	<b>15,954,664</b>	<b>2,746,288</b>
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



Scholarly Data  
oo

Corpus  
oooo●ooo

Artifact Parameters  
oooooooooooo

Misc  
ooooo

# Corpus - Reference Composition

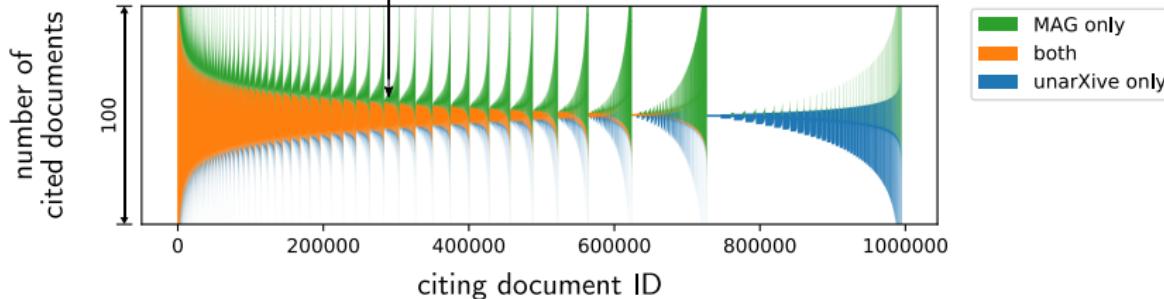
Exemplary visualization  
of a single data point  
on the x-axis.

document #288983

[...]

## References

- [1] Example, A. (1970). Giving an example. *Journal of Examples*, 3, 11-27.
- [2] Füller, F. (2019). Schnibgerie, Zeitschrift der Schreibgeräte, 77, 115-137.
- [3] Beispiel, B. (1990). Beispieldokumentation, 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, R. (2018). The Art of Transcribing Titles. *Journal of Titles*, 1, 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.



Scholarly Data  
oo

Corpus  
oooooo●ooo

Artifact Parameters  
oooooooooooo

Misc  
ooooo

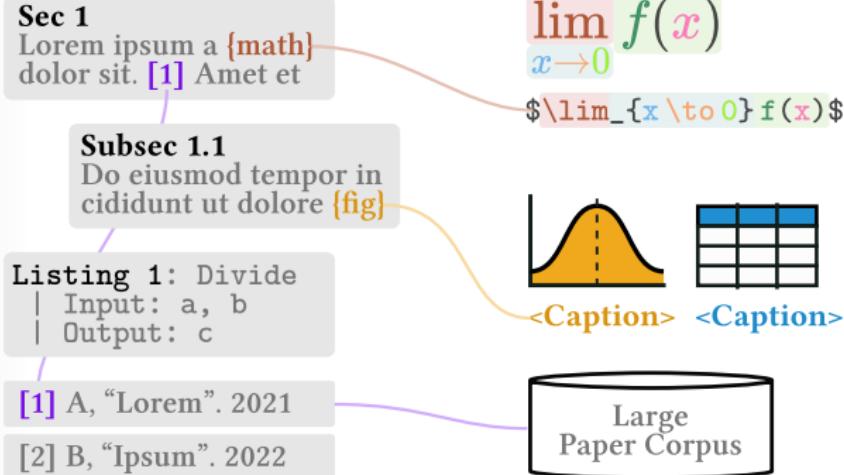
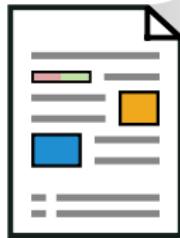
# Corpus - Target Sec. Specific Refs

	Discipline <sup>a</sup>	Count	Normalization factor	Normalized ratio (%)
<b>Citing</b>	Mathematics	298,009	4.66	8.70
	CS	9,123	6.31	0.36
	Physics	30,593	1.72	0.33
<b>Cited</b>	Mathematics	313,651	3.15	6.20
	CS	12,179	8.50	0.65
	Physics	31,087	2.04	0.40
<b>Pairs</b>	<u>Math<sup>†</sup>→Math<sup>‡</sup></u>	200,859	5.41	6.81
	<u>Math<sup>†</sup>→CS</u>	5,134	92.13	2.96
	<u>Math<sup>†</sup>→Phys</u>	3,114	89.88	1.75
	<u>CS→Math<sup>‡</sup></u>	3,456	18.82	0.41
	<u>Phys→Math<sup>‡</sup></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

<sup>a</sup>  $\dagger$ : Mathematics citing document,  $\ddagger$ : Mathematics cited document, X→X: Citing and cited document are from the same discipline.

# unarXive 2022 Schema

LaTeX  
↓  
{json}



- Created from LaTeX sources

- Document structure and content types preserved

- Math, figures, tables, references linked

Scholarly Data  
oo

Corpus  
oooooooo●

Artifact Parameters  
oooooooooooo

Misc  
ooooo

# 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** ( $\alpha$ , 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

Scholarly Data  
oo

Corpus  
oooooooo

Artifact Parameters  
●oooooooooooo

Misc  
ooooo

# 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

Scholarly Data  
oo

Corpus  
oooooooo

Artifact Parameters  
o●oooooooooooo

Misc  
ooooo

# 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 .

Scholarly Data  
oo

Corpus  
oooooooo

Artifact Parameters  
oo●oooooooo

Misc  
oooo

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

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

Scholarly Data  
oo

Corpus  
oooooooo

Artifact Parameters  
oooo●oooooooo

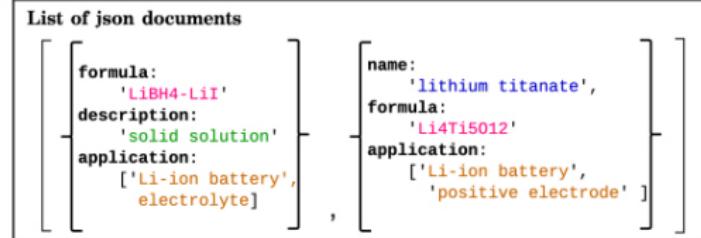
Misc  
oooo

# 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

Scholarly Data  
oo

Corpus  
oooooooo

Artifact Parameters  
oooo●oooooooo

Misc  
oooo

# 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

Scholarly Data  
OO

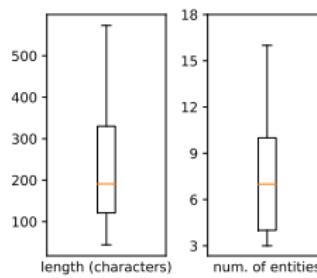
Corpus  
oooooooo

Artifact Parameters  
ooooo●ooooo

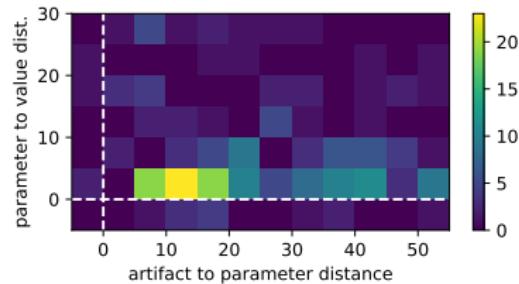
Misc  
ooooo

# 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



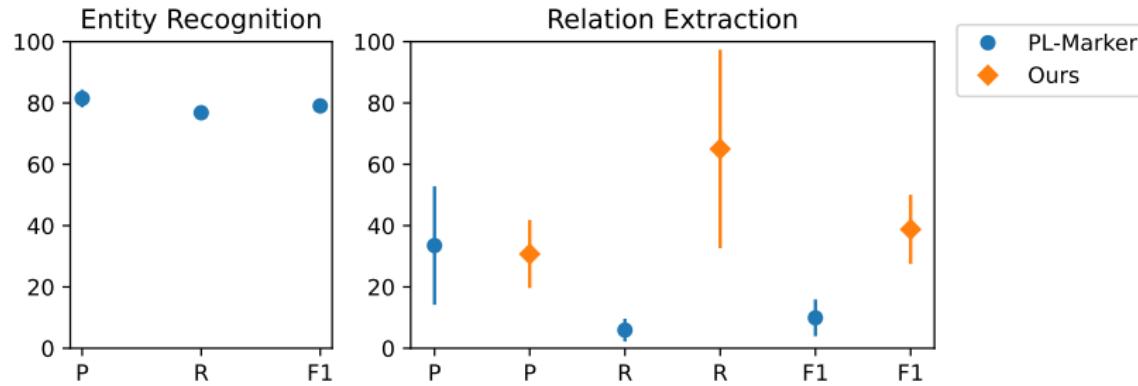
(a) text segments



(b) relation distances (#chars)

Figure: Observations of initial annotation round

# Artifact Parameters - Experiments: Fine-tuned models



Scholarly Data  
oo

Corpus  
oooooooo

Artifact Parameters  
oooooooo●oooo

Misc  
ooooo

# Artifact Parameters - Experiments: Fine-tuned models

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

Model Input	P [%]	R [%]	F <sub>1</sub> [%]
CD	15.5	8.8	11.1
T <sub>CD</sub>	16.6	29.8	19.6
TC <sub>D</sub>	26.5	65.0	35.5
<b>TCD</b>	<b>30.7</b>	<b>65.0</b>	<b>38.8</b>

Scholarly Data  
oo

Corpus  
ooooooooo

Artifact Parameters  
oooooooo●ooo

Misc  
ooooo

# Artifact Parameters - Experiments: LLMs

Zero-shot		Entity Recognition			Relation Extraction		
Model	Output	P [%]	R [%]	F <sub>1</sub> [%]	P [%]	R [%]	F <sub>1</sub> [%]
WizardLM	JSON	6.9	11.3	8.6	0.1	0.8	0.1
	YAML	9.7	35.6	15.3 <sub>△+6.7</sub>	0.1	1.5	0.1 <sub>△+0.0</sub>
Vicuna <sub>4k</sub>	JSON	15.1	9.3	11.5	0.7	3.8	1.2
	YAML	17.3	31.5	22.3 <sub>△+10.8</sub>	0.0	0.8	0.1 <sub>△-1.1</sub>
Falcon	JSON	<b>37.1</b>	5.9	10.2	0.0	0.0	0.0
	YAML	32.7	14.2	19.8 <sub>△+9.6</sub>	0.0	0.0	0.0 <sub>△+0.0</sub>
GALACTICA	JSON	25.9	15.7	19.5	0.1	2.3	0.3
	YAML	23.1	19.5	21.1 <sub>△+1.6</sub>	0.0	0.8	0.1 <sub>△-0.2</sub>
GPT-3.5	JSON	27.9	<b>42.8</b>	33.8	5.4	10.7	7.2
	YAML	<u>34.0</u>	<u>41.7</u>	<u>37.4<sub>△+3.6</sub></u>	<u>5.8</u>	<u>12.2</u>	<u>7.8<sub>△+0.6</sub></u>
5-shot		Entity Recognition			Relation Extraction		
Vicuna <sub>16k</sub>	JSON	34.4	<u>46.7</u>	39.6	0.8	4.6	1.3
	YAML	<b>43.9</b>	<b>44.1</b>	<u>44.0<sub>△+0.4</sub></u>	<u>4.5</u>	<u>9.9</u>	<u>6.1<sub>△+4.8</sub></u>

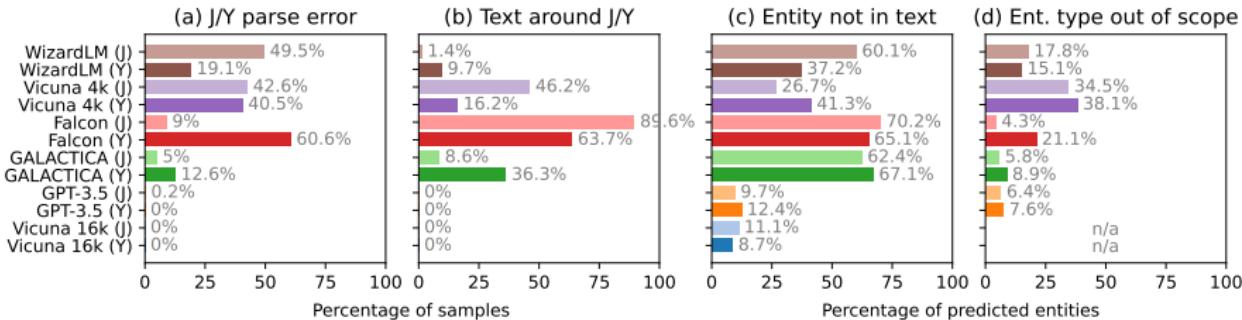
Scholarly Data  
oo

Corpus  
ooooooooo

Artifact Parameters  
oooooooooooo●o

Misc  
ooooo

# Artifact Parameters - Experiments: LLMs



Scholarly Data  
oo

Corpus  
ooooooooo

Artifact Parameters  
ooooooooo●

Misc  
ooooo

# 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 Core<sup>1</sup> (conferences) and SJR<sup>2</sup> (journals).<sup>3</sup>

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

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

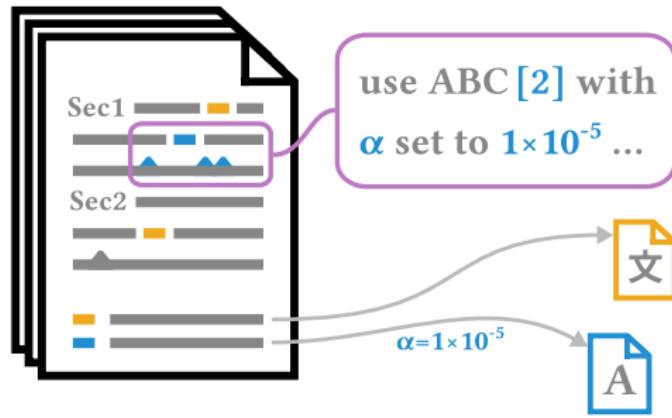
<sup>3</sup>Ratings for publication year or, if not listed, most up-to-date ranking. Workshops ranks are that of the hosting conference.

## 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.

# Quality Dimensions



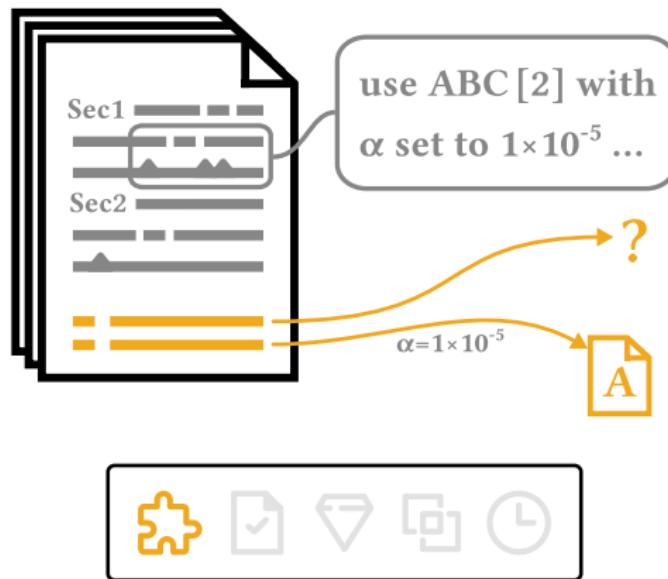
Scholarly Data  
oo

Corpus  
oooooooo

Artifact Parameters  
oooooooooooo

Misc  
oo●oo

# Quality Dimensions: Completeness



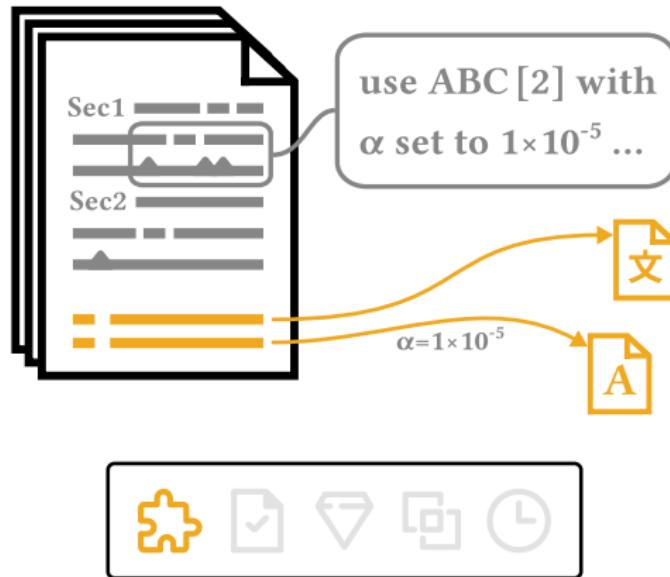
Scholarly Data  
oo

Corpus  
oooooooo

Artifact Parameters  
oooooooooooo

Misc  
oo●oo

# Quality Dimensions: Completeness



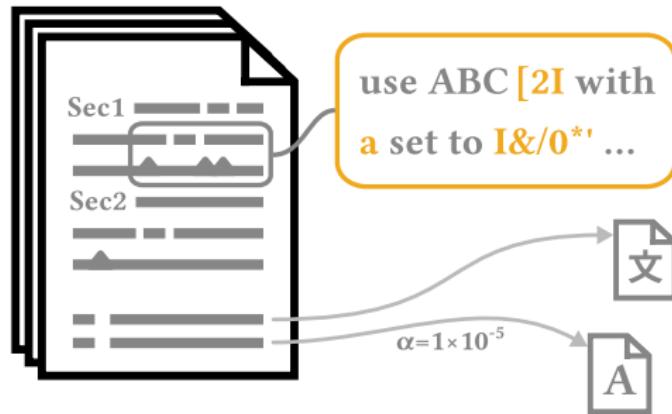
Scholarly Data  
oo

Corpus  
oooooooo

Artifact Parameters  
oooooooooooo

Misc  
oo●oo

# Quality Dimensions: Accuracy



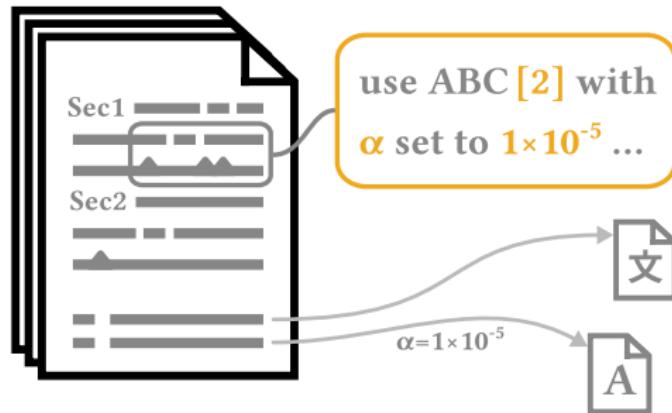
Scholarly Data  
oo

Corpus  
oooooooo

Artifact Parameters  
oooooooooooo

Misc  
oo●oo

# Quality Dimensions: Accuracy



Scholarly Data  
oo

Corpus  
oooooooo

Artifact Parameters  
oooooooooooo

Misc  
oo●oo

# Quality Dimensions: Relevance



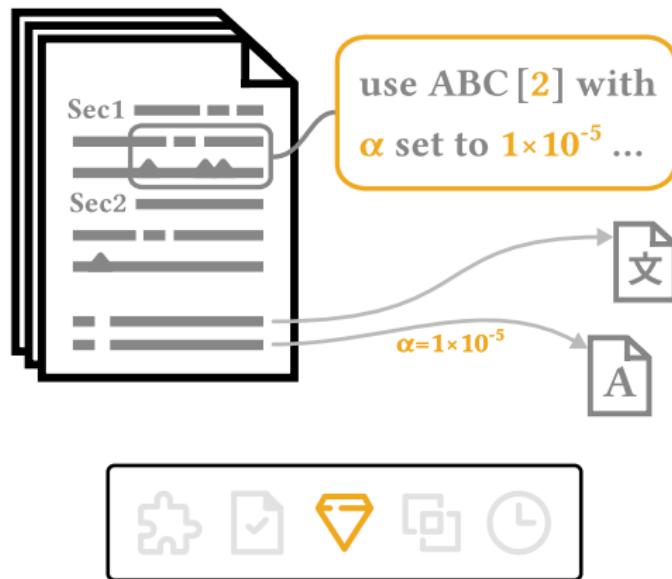
Scholarly Data  
oo

Corpus  
oooooooo

Artifact Parameters  
oooooooooooo

Misc  
oo●oo

# Quality Dimensions: Relevance



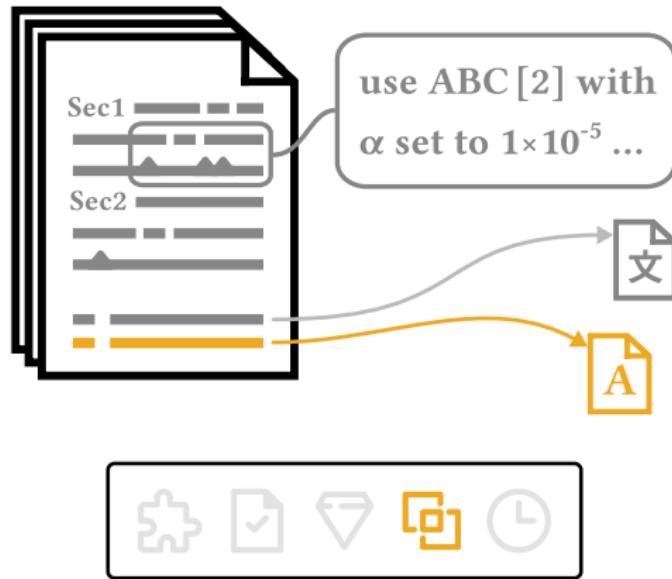
Scholarly Data  
oo

Corpus  
oooooooo

Artifact Parameters  
oooooooooooo

Misc  
oo●oo

# Quality Dimensions: Comparability



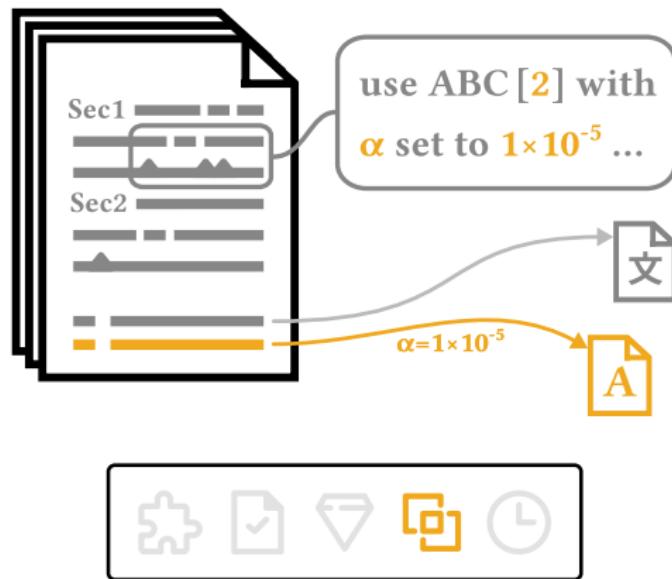
Scholarly Data  
oo

Corpus  
oooooooo

Artifact Parameters  
oooooooooooo

Misc  
oo●oo

# Quality Dimensions: Comparability



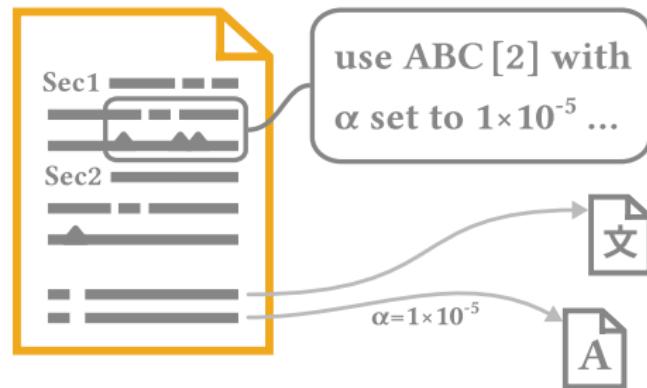
Scholarly Data  
oo

Corpus  
oooooooo

Artifact Parameters  
oooooooooooo

Misc  
oo●oo

# Quality Dimensions: Timeliness



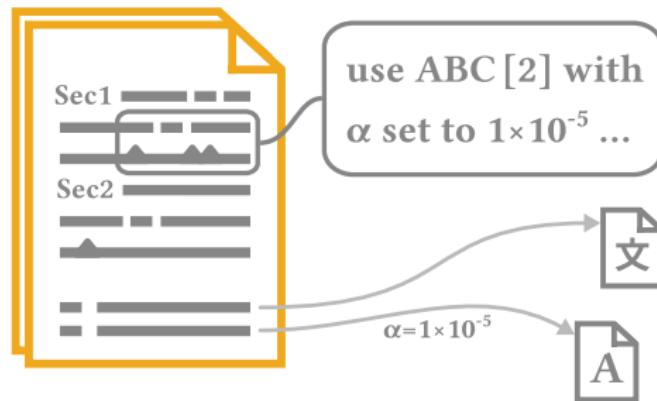
Scholarly Data  
oo

Corpus  
oooooooo

Artifact Parameters  
oooooooooooo

Misc  
oo●oo

# Quality Dimensions: Timeliness



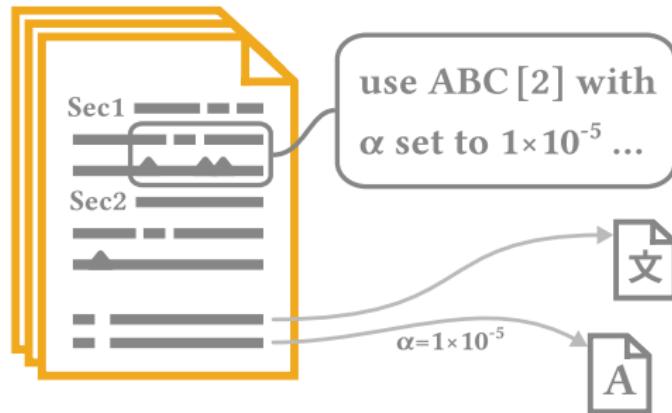
Scholarly Data  
oo

Corpus  
oooooooo

Artifact Parameters  
oooooooooooo

Misc  
oo●oo

# Quality Dimensions: Timeliness



Scholarly Data  
oo

Corpus  
oooooooo

Artifact Parameters  
oooooooooooo

Misc  
oo●oo

# Hardest Challenges

- general: **Reference Matching**

solved how:

- solid base understanding/approach
  - title+author fuzzy matching / blocking
  - ref. parser & target set choice
- domain knowledge: APS journal DOI determinability
- model training: reference parser training on synthetic training data

- most underestimated / labor intensive: **HyperPIE data set**

details:

- conceptual: data scheme, annotation guidelines
- technical: tool (offsets (UI vs backend, newlines))

- most “finicky”: **RE evaluation** with partial matches on entity level

Scholarly Data  
oo

Corpus  
oooooooo

Artifact Parameters  
oooooooooooo

Misc  
○○●○

# Limitations

- **Corpus**

- L<sup>A</sup>T<sub>E</sub>X 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

Scholarly Data  
oo

Corpus  
oooooooo

Artifact Parameters  
oooooooooooo

Misc  
oooo●