Harvesting Textual and Structured Data from the HAL Publication Repository

Francis Kulumba, Wissam Antoun, Guillaume Vimont, Laurent Romary

▶ To cite this version:

Francis Kulumba, Wissam Antoun, Guillaume Vimont, Laurent Romary. Harvesting Textual and Structured Data from the HAL Publication Repository. 2024. hal-04665456

HAL Id: hal-04665456
https://hal.science/hal-04665456

Preprint submitted on 31 Jul 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Harvesting Textual and Structured Data from the HAL Publication Repository

Francis Kulumba[1, 2] **Wissam Antoun**[1,2]
Guillaume Vimont[3] **Laurent Romary**[1]
[1]Inria, Paris, France [2]Sorbonne Université, Paris, France [3]IRIF, Paris, France
{francis.kulumba, wissam.antoun, laurent.romary}@inria.fr
vimont@irif.fr

Abstract

HAL (*Hyper Articles en Ligne*) is the French national publication repository, used by most higher education and research organizations for their open science policy. As a digital library, it is a rich repository of scholarly documents, but its potential for advanced research has been underutilized. We present HALvest, a unique dataset that bridges the gap between citation networks and the full text of papers submitted on HAL. We craft our dataset by filtering HAL for scholarly publications, resulting in approximately 700,000 documents, spanning 56 languages across 13 identified domains, suitable for language model training, and yielding approximately 16.5 billion tokens (with 8 billion in French and 7 billion in English, the most represented languages). We transform the metadata of each paper into a citation network, producing a directed heterogeneous graph. This graph includes uniquely identified authors on HAL, as well as all open submitted papers, and their citations. We provide a baseline for authorship attribution using the dataset, implement a range of state-of-the-art models in graph representation learning for link prediction, and discuss the usefulness of our generated knowledge graph structure.

1 Introduction

Publication repositories are the norm when storing and distributing scholarly papers openly in a sovereign and sustainable way. When submitting a paper to a repository, the depositor, an identified user, possesses a unique `id` in the repository's database. However, the depositor's co-authors might not be identified users, letting the repository perform `id` retrieval to map the submission to the correct authors. Authorship attribution without human interventions remains a non-trivial task to publication repositories (Tekles and Bornmann, 2019). Researchers rapidly apprized the importance of co-authorship to tackle authorship attribution, hence

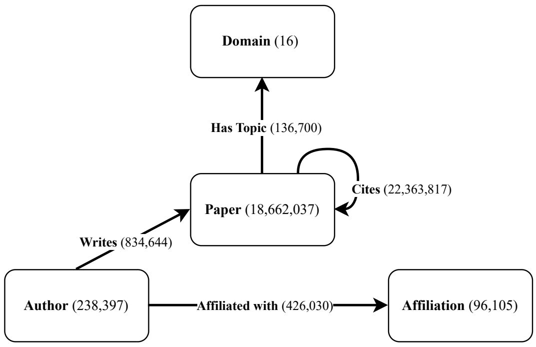

Figure 1: HALvest's citation network: a directed heterogeneous graph with 4 node types and 4 edge types.

integrated structured data to their proposed solutions (Shin et al., 2014; Ma et al., 2020; Xie et al., 2022). However, the shortcomings of graph-based algorithms, in the absence of co-authorship, drove researchers to exploit semantic and symbolic features from papers (Tran et al., 2014; Müller, 2017; Kim et al., 2019; Cohan et al., 2020; Boukhers et al., 2021; Pooja et al., 2021; Kojaku et al., 2021) and investigating the semantic fingerprint from individuals (Han et al., 2017; Zhai et al., 2019).

Recently, with the research in multimodal deep-learning mushrooming, several architectures allow practitioners to now benefit from text and structured data (Zhang et al., 2019, 2022; Yasunaga et al., 2022), thus exploiting citation networks as well as semantic content from scholarly papers (Pooja et al., 2022; Santini et al., 2022). By opening the data from Hyper Articles en Ligne (HAL), we aim to enable researchers to implement and validate new methods towards better multi-modal architectures—exploiting graphs and text—and in an ancillary manner tackling authorship identification. We introduce HALvest, a structured and textual dataset: the structured part, called HALvest-Geometric, is a heterogeneous citation network comprising 238,397 author nodes and 18,662,037

Dataset	References	Full-text	Multilingual	Multi-domain
		CURATED GRAPH DATASETS		
ArnetMiner (Tang et al., 2012)	✓			
		LARGE GRAPH DATASETS		
MAG (Wang et al., 2020)	✓			✓
OpenAlex (Priem et al., 2022)	✓	✓	✓	✓
		OURS		
HALvest	✓	✓	✓	✓

Table 1: Comparison of information provided by HALvest to previous academic graph datasets. References denote the presence of a citation list for a given paper.

paper nodes, with 642,723 of them having their full-text available. Furthermore, the gathered papers allow more than 16.5 billion tokens from scholarly text, allowing for better language modeling in a given domain.

To validate HALvest's suitability to not only train language models but also graph representation learning architectures, we conduct a straightforward experiment: we test several state-of-the-art graph neural networks (Scarselli et al., 2009), that can fit heterogeneous graphs, to provide a baseline for authorship attribution in a closed set-up, while also validating our graph's format.

Our contribution can be summarized as follows

- A textual dataset comprising 17 billion tokens in 56 languages and 13 domains.
- An academic citation network with 238,397 disambiguated authors and 18,662,037 scholarly papers.

The datasets [1] [2], as well as the code used to craft them [3] [4] are available online and will be actively maintained and updated.

2 Related Work

Scholarly textual data The rise in popularity of language modeling in natural language processing, coupled with online publication repositories being an integral part of the researcher's toolkit, allowed the release of a non-negligible amount of textual data. Although full-texts are not bound to the same licensing as abstracts, publication libraries in science, technology, engineering, and math are often at the forefront when it comes to opening their data when possible. The likes of ArXiv [5] and DBLP [6] PubMed [7] are allowing their publication's titles and abstract to be crawled (arXiv.org submitters, 2024; Wahle et al., 2022; Sen et al., 2008; Doğan et al., 2014) while Semantic Scholar [8] and a subset of PubMed (PMC) [9], allow for a full-texts' crawling (Ammar et al., 2018; Lee et al., 2020) of their data. ACL [10], on the other hand, opted for a cleaned subset of 10,920 academic papers from ACL Anthology (Bird et al., 2008). However, the aforementioned repositories are either limited in size or domains (e.g., STEM for ArXiv, computer science for DBLP, and bio-medical for PubMed), hence Lo et al. (2020) merged various sources in order to alleviate these issues, while also offering full-texts from open some papers. Academic unstructured data is also built and maintained by researchers' communities as seen in BC5CDR (Li et al., 2016) and NCBI (Doğan et al., 2014), and MEDLINE (Vishnyakova et al., 2019) in the biomedical domain. All the aforementioned datasets are confined to the English language and featured publications or preprints, while the unstructured half of HALvest provides historical texts, thesis, and dictionaries, for example, being more lengthy and domain-diverse. Besides, the full texts are uniform in their formatting and referencing style.

Academic graph When it comes to academic graph, ArnetMiner (Tang et al., 2008, 2012) and Microsoft Academic Graph (MAG)(Wang et al., 2020) are often the go-to, comprising of meta-

[1] https://huggingface.co/datasets/almanach/HALvest

[2] https://huggingface.co/datasets/Madjakul/HALvest-Geometric

[3] https://github.com/Madjakul/HALvesting

[4] https://github.com/Madjakul/HALvesting-Geometric

[5] https://arxiv.org/

[6] https://dblp.org/

[7] https://pubmed.ncbi.nlm.nih.gov/download/

[8] https://www.semanticscholar.org/

[9] https://www.ncbi.nlm.nih.gov/pmc/

[10] https://www.aclweb.org/portal/

data from DBLP and ACM [11]. MAG, unlike ArnetMiner, has not been manually curated but allows for a rich pool of papers and authors (736,389 papers, 1,134,649 authors 8,740 institutions, and 59,965 fields of study nodes). One can also look upon OpenAlex (Priem et al., 2022), offering more than 220 million publication nodes, and the graph proposed by (Ammar et al., 2018) to enable semantic features in Semantic Scholar. All the graphs mentioned above come with various metadata, including abstracts.

Combined Text and academic knowledge graph To the best of our knowledge, the only ready-to-use academic graph dataset mapped with full texts is from OpenAlex (Priem et al., 2022). However, ACL (Cohan et al., 2019), and Lee et al. (2020)'s datasets can be easily mapped to a citation network with little processing. Despite providing the full text of every paper's node, ACL and Lee et al. (2020) graphs are limited in size and domain (computational linguistic and biomedical respectively).

3 Collection Methodology

Crafting HALvest is a four-step process. We start by fetching open PDF files from HAL, before using GROBID (GROBID Repository) to obtain XML files and easing the conversion to TXT documents. We filter the French and English TXT files out of poorly encoded text, and use the remaining documents to build the citation network. In the following, we introduce our pipeline in more detail.

3.1 Fetching data from HAL

HAL's API allows for precise requests to the repository. One can acquire structured XML-TEI responses of each submitted document. Exploiting XML-TEIs allows for standardized access to all the metadata available on HAL, and therefore, more control over the desired document representation. We settled for a list of eight features to represent submissions:

- `halid`: submission's unique identifier assigned by HAL.
- `lang`: language of the document, as filled by the depositor.
- `title` title of the document.
- `domain`: list of field of studies [12].
- `timestamp`: time of access.
- `year`: publication year of the document if relevant. Otherwise, it is set to year 1.
- `url`: URL to access the PDF.
- `authors`: list of authors.

The `author` is comprised of at most nine features:

- `name`: string for the author name, as filled by the depositor.
- `affiliations`: list of unique identifiers attributed by HAL to the institutions where the author belongs.
- `halauthorid`: unique identifier assigned by HAL to each author registered on the online repository. If an author is not registered to HAL, he is considered unidentified and is assigned an `halauthorid` of "0".
- Potentially six external identifiers, if provided by the registered authors, comprising of `arxiv`, `researcherid`, `idref`, `orcid`, `viaf` and `isni`.

We design our query to only fetch submissions with an attached document in open access, and apply further filtering to only fetch work submitted alongside a PDF file. As of February 2024, the total number of open submitted PDF files is 778,072.

We use GROBID (GROBID Repository) to derive XML files from PDFs, hence easing the conversion to plain text afterward. Reference markers in each document are serialized following the scheme defined by Taylor et al. (2022). However, math, SMILES formulas, as well as code, are encoded in plain utf-8 and left as is.

3.2 Filtering

During the conversion process, if the fonts within a PDF lack Unicode tables and do not employ standard encoding for mapping glyph indices to characters, GROBID's output ends up being gibberish (§ A.2), as the latter do not employ optical character recognition to extract text from PDFs, but rather its layout. To filter out gibberish documents, we use a set of heuristic functions. Following Raffel et al. (2020); Wenzek et al. (2020); Rae et al. (2022) and Penedo et al. (2023), we compute a set of statistics about each document, effectively getting metrics

[11] https://dl.acm.org/
[12] https://hal.science/browse/domain

like the number of lines, the average word length or the ratio of unique words. We use the implementation provided by RedPajama Repository. To compute the ratio of stop-words in a document, we use `stopwords-json` (6, 2024).

We post-process the remaining documents, written in 34 languages, as follows :

1. Documents with less than 3 words are discarded.
2. Documents with more than 10% of words that are capitalized are discarded.
3. Documents with more than 60% of words that are not alphanumeric are discarded.
4. Documents with an average word length of 1.5 characters or below are discarded.
5. Document with no stop words are discarded. Stop words are strong indicators for well-redacted documents, hence, lowering the chance of it being gibberish. Besides, we use stop words as a language identifier, as the language provided in the metadata is specified by a depositor, and human error can be introduced.
6. We compute the inverse fertility (Rust et al., 2021): the number of words in a document divided by the number of tokens yielded by the mT5 tokenizer (Xue et al., 2021). After removing special tokens, a tokenizer yields at most an amount of tokens equal to the number of words, effectively bounding our function between 0 and 1. An inverse fertility score close to zero is a strong indication of over tokenization, and therefore, hinting at a gibberish document. Documents with an inverse fertility score lower than 0.2 are discarded.

3.3 From metadata to citation network

Building the citation network is straightforward, though, computationally demanding. We define four node types, and four edge types, before iterating through the metadata of each document. Proceeding this way leads to several shortcomings: the computed graph represents a snapshot of HAL at a given time, remaining static, leading to affiliations being cumulative—affiliation nodes do not represent the affiliations of an author at a given time, but rather all the institutions he has been part of. To alleviate those issues, we provide optional features for the author and paper nodes, allowing practitioners to only compute edges for a subset of the graph, based on a year or a language.

When it comes to domain identification, HAL provides a tree-like domain space, with 13 domains, branching into several subdomains, and an additional one: "other". We fetch the domains filled by the depositors and use the root node of each one. We obtained 16 domains, as some depositors only provide a subdomain that overlaps between two parent domains.

Regardless of how we compensate for the lack of precision in some of the graph's relations, computing the citations remains fuzzy. We retrieve each document's references through their XML document, collect their title and publication year, and then use exact matching. However, this process can induce inaccuracies in the title retrieved, as GROBID parsing can be inexact.

4 Dataset Composition

In the following section, we provide more details concerning the composition of the two halves of HALvest: the unstructured part with text, and the structured part with a heterogeneous citation network.

4.1 Unstructured data

Although HALvest is mostly in English and French, the gathered 670,861 papers are written in 56 languages across 16 domains for the unfiltered version, accounting for approximately 17 billion tokens. HALvest's text can also serve as a valuable asset for low-resource languages, hosting documents in Basque, Catalan, or Persian to mention a few (§B.1).

4.2 Structured Data

HALvest-Geometric is made up of a heterogeneous graph. Following Wang et al. (2020), we compute 238,397 author nodes, 18,662,037 paper nodes, 96,105 institution nodes, and 16 domain nodes for a total of18,996,55 nodes as of February 2024. We also define 4 different edge types, constituting 23,761,191 edges.

4.3 Do citations help, even when fuzzy?

In this subsection, we will evaluate the citations extracted from each publication, as described in subsection 3.3. Because we fuzzily retrieve the citations, the added information to the citation network can be nothing more than noise. A straight-

Key	Value
halid	01744328
lang	en
domain	["info.info-ai"]
timestamp	2024/03/05 22:32:07
year	2017
url	https://hal.science/hal-01744328/file/ertek_chi_zhang_2017_RFID.pdf
text	A Framework for Mining RFID Data From Schedule-Based Systems Gürdal Ertek, Xu Chi, Member, IEEE, and Allan N. Zhang Member, IEEE Abstract-A schedule-based system is a system that operates on or contains within a schedule of events and breaks at particular time intervals...

Table 2: Examples of text documents from HALph.

forward way to estimate the usefulness of this retrieved information is to use graph neural networks (GNN) and message passing. Message passing allows GNNs to discover the graph's structure from the way information propagates on it. Another key concept to better understand our experiment is a property called homophily: adjacent nodes should have similar features; in our context since authors would most likely cite people from their fields, the assumption that our data is highly homophilic is not far-fetched. Therefore, the subsequent graph neural networks used to represent our graph should perform better in setting with domain-related papers as adjacent nodes, than with other papers two hopes away, bridged by a domain node.

4.3.1 Tasks & Settings

Authorship attribution Given a document $s \in \mathcal{S}$ and a candidate author $a \in \mathcal{A}$, we want to compute a probability

$$\begin{aligned} p : \mathcal{S} \times \mathcal{A} &\mapsto [0, 1] \\ s, a &\rightarrow p(y|s, a) \end{aligned}$$

that the individual a is an author of s.

Link prediction When given a citation network, the authorship attribution problem can be reformulated as a link prediction problem. Given a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ with V the set of all nodes and $\mathcal{E} \subseteq |\mathcal{V}| \times |\mathcal{V}|$ the set of all the true edges in $\mathcal{G}$. In our setting, given a potential author node u and a paper node v, we want to learn a classifier that predicts the probability of the existence of an edge, by computing a score between the representations of both incident nodes $\hat{y}_{u\sim v} = f(h_u, h_v)$ where h_u and h_v are learned representation of said nodes. In this task we use a set of candidate edges $\mathcal{E}'$ comprising of positive and negative edges between the incident nodes, allowing us to use a binary cross entropy loss

$$\begin{aligned} \mathcal{L} = -\sum_{u\sim v \in E'} &[y_{u\sim v} \log(\hat{y}_{u\sim v}) \\ &+ (1 - y_{u\sim v}) \log(1 - \hat{y}_{u\sim v})] \end{aligned}$$

This closed setting, however, doesn't account for name ambiguity and unknown authors, as we only consider identified candidates from the citation network to perform link prediction.

Baselines HALvest's heterogeneous graph is featureless, prompting us to learn a representation for each node. We use several state-of-the-art graph neural networks (GNN) architectures (Scarselli et al., 2009) to learn embedding for each node while training models.

Evaluation We report the area under the curve (AUC) as a measure of the quality of a link prediction algorithm. Since our experiment is done in a closed setting, we compute five random splits of the original citation network, keeping 10% of the author $\leftrightarrow$ paper edges for validation and 20% for test purposes.

4.3.2 Link Prediction

We build a simple link prediction architecture, comprising of an embedding layer for each node type—of dimension $\mathbb{R}^{|\mathcal{V}'| \times 16}$ where $|\mathcal{V}'|$ is the number of nodes of a given type—followed by two GNN layers, to learn an inductive 16 dimensions representation of each node. The implemented GNNs are GraphSAGE (Hamilton et al., 2017), a graph attention network (Veličković et al., 2018), and a residual gated graph convolutional network (Bresson and Laurent, 2018). Finally, we compute the

GNN	AUC	AUC (w/o citations)
GraphSAGE	$99.08_{\pm 0.05}$	$89.56_{\pm 0.13}$
GAT	$98.44_{\pm 0.09}$	$74.30_{\pm 5.10}$
RGGC	$99.3_{\pm 0.05}$	$91.71_{\pm 0.29}$

Table 3: Area under the curve (with standard deviation) of each link prediction model. The result is the average performance of 5 models trained with 5 different random seeds. We use GraphSage (Hamilton et al., 2017), graph attention network (Veličković et al., 2018) and residual gated graph convolutions (Bresson and Laurent, 2018).

cosine similarity between candidate author and paper nodes, map it to probabilities, and rule out, or not the presence of a link between each pair of nodes.

We follow (Fey and Lenssen, 2019)'s implementation and perform Bayesian optimization to find the best hyperparameters (Li et al., 2020; Bergstra et al., 2015). We report the training results in Table 3. The training hyperparameters and learning metrics can be found in § C.1.

4.3.3 Results

The results of our experiments provide compelling evidence that incorporating citations, even when retrieved fuzzily, enhances the performance of graph-based models in authorship attribution tasks. The area under the curve (AUC) scores, as presented in Table 3, illustrate this improvement across various graph neural network (GNN) architectures.

Impact of Citations on Performance Our evaluation metric, clearly indicates that the inclusion of citation information significantly boosts the model's performance by about 10% across all models. These results underscore the hypothesis that citations, even when not perfectly precise, contribute valuable information to the citation network. This enhancement can be attributed to the property of homophily, where nodes (authors and papers) tend to be more similar when they are closely related or cited within the same domain.

Model Comparisons and Robustness The comparative analysis of different GNN architectures further validates our approach. All three models—GraphSAGE, GAT, and RGGC—exhibited significant performance drops when citation information was removed. This consistency across multiple architectures suggests that the benefits of incorporating citations are not model-specific but rather a generalizable advantage. Moreover, the standard deviations reported alongside the AUC scores reflect the robustness of our models. The relatively low standard deviations indicate stable performance across different random splits of the data, reinforcing the reliability of our findings.

5 Limitations

PDF processing Solely relying on GROBID to process the PDFs hindered the dataset creation in some aspects. As only the PDF's layout tokens matter to GROBID, and no optical character recognition is performed, documents with odd layouts can not be exploited after being converted, and are outright discarded. Most of the documents are seamlessly converted, yet some of them have near gibberish span of text within cleaner sentences. We designed our filtering process to remove documents with a prior unusable layout, but not to remove the gibberish span of text within clean data. Therefore, further processing effort is needed from practitioners.

Multilingual documents Some documents are written in two or more languages, but HAL imposes the depositor to only fill one language. It is therefore necessary to perform language identification at the sentence level to ensure only the needed language is fetched from the dataset.

Accounting for references When building the citation network, the references at the end of each paper are not consolidated, resulting in a fuzzy process of deduplication afterward. This leads our graph to lack correctness when it comes to modeling citations.

6 Conclusion

By processing approximately 700,000 documents in 56 languages across 13 domains from Hyper Articles en Ligne (HAL), we have created a unique resource that maps natural language text to a directed heterogeneous graph: HALvest, a comprehensive dataset that integrates citation networks with the full-text of scholarly papers. This dataset encompasses 16.5 billion textual tokens for 18,996,55 nodes, enabling extensive research in multimodality, authorship attribution, author name disambiguation, domain classification, and more.

We elaborate on the usefulness of HALvest, by performing authorship attribution, using state-of-the-art GNN architectures, and discuss the added value of retrieved citations, further confirming

HALvest's adequacy for the natural language processing and graph representation learning fields.

Future work will focus on expanding the dataset and improving the preprocessing pipeline, for greater accuracy representing the academic interactions, and increased utility.

The dataset, along with associated benchmarks, are openly available, aiming to democratize access to large-scale scholarly data.

Acknowledgements

The authors are grateful to the CCSD staff, Achraf Azhar, and Patrice Lopez for providing resources and support to better handle HAL and GROBID. We would also like to thank Brahim Talb and Menel Mahamdi for the productive discussions.

References

6. 2024. stopwords-json. Original-date: 2014-02-01T08:08:26Z.

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. 2018. Construction of the Literature Graph in Semantic Scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 84–91, New Orleans - Louisiana. Association for Computational Linguistics.

arXiv.org submitters. 2024. arxiv dataset.

James Bergstra, Brent Komer, Chris Eliasmith, Dan Yamins, and David D Cox. 2015. Hyperopt: a python library for model selection and hyperparameter optimization. *Computational Science & Discovery*, 8(1):014008.

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

Zeyd Boukhers, Nagaraj Bahubali, Abinaya Thulsi Chandrasekaran, Adarsh Anand, Soniya Manchenahalli Gnanendra Prasadand, and Sriram Aralappa. 2021. Bib2Auth: Deep Learning Approach for Author Disambiguation using Bibliographic Data. *arXiv preprint*. ArXiv:2107.04382 [cs].

Xavier Bresson and Thomas Laurent. 2018. Residual Gated Graph ConvNets. *arXiv preprint*. ArXiv:1711.07553 [cs, stat].

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural Scaffolds for Citation Intent Classification in Scientific Publications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3586–3596, Minneapolis, Minnesota. Association for Computational Linguistics.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

Matthias Fey and Jan E. Lenssen. 2019. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.

GROBID Repository. 2008–2024. Grobid. `https://github.com/kermitt2/grobid`. *Preprint*, swh:1:dir:dab86b296e3c3216e2241968f0d63b68e8209d3c.

Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Hongqi Han, Changqing Yao, Yuan Fu, Yongsheng Yu, Yunliang Zhang, and Shuo Xu. 2017. Semantic fingerprints-based author name disambiguation in Chinese documents. *Scientometrics*, 111(3):1879–1896.

Kunho Kim, Shaurya Rohatgi, and C. Lee Giles. 2019. Hybrid Deep Pairwise Classification for Author Name Disambiguation. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, pages 2369–2372, New York, NY, USA. Association for Computing Machinery.

Sadamori Kojaku, Xiaoran Yan, Jisung Yoon, Filipi N. Silva, Vincent Larivière, and Yong-Yeol Ahn. 2021. DisamBERT: Author name disambiguation with BERT.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068.

Liam Li, Kevin Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Jonathan Ben-tzur, Moritz Hardt, Benjamin Recht, and Ameet Talwalkar. 2020. A system for massively parallel hyperparameter tuning. In *Proceedings of Machine Learning and Systems*, volume 2, pages 230–246.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Yingying Ma, Youlong Wu, and Chengqiang Lu. 2020. A Graph-Based Author Name Disambiguation Method and Analysis via Information Theory. *Entropy*, 22(4):416.

Mark-Christoph Müller. 2017. Semantic Author Name Disambiguation with Word Embeddings. In *Research and Advanced Technology for Digital Libraries*, Lecture Notes in Computer Science, pages 300–311, Cham. Springer International Publishing.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data Only. *Advances in Neural Information Processing Systems*, 36:79155–79172.

K. M. Pooja, Samrat Mondal, and Joydeep Chandra. 2022. Online author name disambiguation in evolving digital library. *Neurocomputing*, 493:1–14.

KM. Pooja, Samrat Mondal, and Joydeep Chandra. 2021. Exploiting similarities across multiple dimensions for author name disambiguation. *Scientometrics*, 126(9):7525–7560.

Jason Priem, Heather Piwowar, and Richard Orr. 2022. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint*. ArXiv:2205.01833 [cs].

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. *arXiv preprint*. ArXiv:2112.11446 [cs].

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):140:5485–140:5551.

RedPajama Repository. 2023. RedPajama: an Open Dataset for Training Large Language Models.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Cristian Santini, Genet Asefa Gesese, Silvio Peroni, Aldo Gangemi, Harald Sack, and Mehwish Alam. 2022. A knowledge graph embeddings based approach for author name disambiguation using literals. *Scientometrics*, 127(8):4887–4912.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2009. The Graph Neural Network Model. *IEEE Transactions on Neural Networks*, 20(1):61–80. Conference Name: IEEE Transactions on Neural Networks.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. 2008. Collective Classification in Network Data. *AI Magazine*, 29(3):93–93. Number: 3.

Dongwook Shin, Taehwan Kim, Joongmin Choi, and Jungsun Kim. 2014. Author name disambiguation using a graph model with node splitting and merging based on bibliographic information. *Scientometrics*, 100(1):15–50.

Jie Tang, Alvis C.M. Fong, Bo Wang, and Jing Zhang. 2012. A Unified Probabilistic Framework for Name Disambiguation in Digital Library. *IEEE Transactions on Knowledge and Data Engineering*, 24(6):975–987. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. ArnetMiner: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 990–998, New York, NY, USA. Association for Computing Machinery.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A Large Language Model for Science. *arXiv preprint*. ArXiv:2211.09085 [cs, stat].

Alexander Tekles and Lutz Bornmann. 2019. Author name disambiguation of bibliometric data: A comparison of several unsupervised approaches. *arXiv preprint*. ArXiv:1904.12746 [cs].

Hung Nghiep Tran, Tin Huynh, and Tien Do. 2014. Author Name Disambiguation by Using Deep Neural Network. In *Intelligent Information and Database Systems*, Lecture Notes in Computer Science, pages 123–132, Cham. Springer International Publishing.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations*.

Dina Vishnyakova, Raul Rodriguez-Esteban, and Fabio Rinaldi. 2019. A new approach and gold standard toward author disambiguation in MEDLINE. *Journal of the American Medical Informatics Association : JAMIA*, 26(10):1037–1045.

Jan Philip Wahle, Terry Ruas, Saif Mohammad, and Bela Gipp. 2022. D3: A Massive Dataset of Scholarly Metadata for Analyzing the State of Computer Science Research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2642–2651, Marseille, France. European Language Resources Association.

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

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Wenjin Xie, Siyuan Liu, Xiaomeng Wang, and Tao Jia. 2022. Author Name Disambiguation via Heterogeneous Network Embedding from Structural and Semantic Perspectives. In *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 245–250. ArXiv:2212.12715 [cs].

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D. Manning, Percy S. Liang, and Jure Leskovec. 2022. Deep Bidirectional Language-Knowledge Graph Pretraining. *Advances in Neural Information Processing Systems*, 35:37309–37323.

Xiaorui Zhai, Hongqi Han, Zhong Li, and Yaxin Ran. 2019. Research on Author Name Disambiguation Based on Fusion Features and Semantic Fingerprints. *Journal of Physics: Conference Series*, 1302(2):022013. Publisher: IOP Publishing.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D. Manning, and Jure Leskovec. 2022. GreaseLM: Graph REASoning Enhanced Language Models for Question Answering. *arXiv preprint*. ArXiv:2201.08860 [cs].

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

A Collection Methodology

A.1 API request format

We use the following request to fetch open papers from HAL, yielding 743,160 documents: https://api.archives-ouvertes.fr/search/?q=*&fq=dateLastIndexed_tdate:[2001-01-01T00:00:00ZTO2024-02-29T23:59:59Z]&fq=openAccess_bool:true&wt=xml-tei&sort=docidasc&rows=500&cursorMark=*

A.2 Example of texts from PDF files with faulty unicode mappings

halid	Sample Text
01762182	1 Introduction 1 IFI wodel nd prolem sttement F F F F F F F F F F F F F F F F F F I IFIFI ystem model F F F F F F F F F F F F F F F F F F F F F F F F I IFIFP vrge dimensionl regime F F F F F F F F F...
00177057	, 4 @ / A -@ & $ B 4 & A / 2 @ - " 0 / -4 , / 3 $C" ' 5$ B 5C % " +0 2 C % BB * (($ - $ @" ' + 0 D 0 E $ F $&$@" C % < = $ B 7 (#7 $ - $ @" 0 D 0 E @" @ C G $ (> ((< 5HI 97>89 D (0 E " @C"@F C G $...
01770410	ACKNOWLEDGMENT S T h i s d o c t o r a l w o r k i s a b o u t m a k i n g o u r s w h a t w e s h a r e . Y e t , I c o u l d n o t h a v e r e a c h e d t h i s f i n a l s t a g e w i t h o u t s h...
01784066	II Theory L I S T O F T A B L E S Table Deterministic exact approaches to mean-variance portfolio selection problem (see also [START_REF] Mansini | Twenty Years of Linear Programming Based Portfolio Optimization[END_REF]...

Table 4: Examples of halid and output text from PDF files with no Unicode mapping, leading to gibberish text. While the first two examples are from discarded documents, the third one is kept, as only the section names are not processed correctly.

See Table 4.

B Composition

B.1 Language composition

See Tables 5 and 6.

B.2 Domain composition

See Table 7.

C Experiments

We train all the models for 2 epochs, with a batch size of 128. For message passing, we sample 32 random neighbor nodes in the first hope and 16 nodes in the second hope. We chose those numbers because the average number of citations is 20, sampling 32 neighboring nodes from an author $\leftrightarrow$ paper edge allows us to capture most of the cited papers as well as the authors. For each positive author $\leftrightarrow$ paper, we sample two negative pairs for the model to train on, those pairs are sampled at each step.

C.1 Link prediction hyperparameters

We used the following hyperparameters for each moedl:

- GraphSage:
  - Hidden channels: 64
  - Dropout: 0.1
  - Weight Decay: 1×10^{-7}
  - Learning rate: 1×10^{-3}
- GAT:
  - Hidden channels: 64
  - Dropout: 0.5
  - Number of attention heads: 8
  - Weight Decay: 1×10^{-4}
  - Learning rate: 5×10^{-3}
- RGGC:
  - Hidden channels: 16
  - Dropout: 0.1
  - Weight Decay: 1×10^{-4}
  - Learning rate: 1×10^{-2}

ISO-639	Language	# Documents	# mT5 Tokens
en	English	464,679	8,158,933,235
fr	French	199,216	9,018,529,985
es	Spanish	2,975	69,221,667
it	Italian	1,172	48,747,986
pt	Portuguese	934	32,918,832
de	German	652	12,225,960
ru	Russian	245	5,763,532
zh	Chinese	160	2,861,585
eu	Basque	113	2,297,485
ar	Arabic	92	2,167,431
ja	Japanese	92	547,861
el	Greek	54	1,738,878
pl	Polish	43	987,878
ro	Romanian	39	1,298,901
uk	Ukrainian	34	837,793
vi	Vietnamese	29	436,660
ca	Catalan	28	975,078
da	Danish	27	961,955
oc	Occitan	26	285,334
br	Breton	24	998,088
sr	Serbian	24	336,878
ko	Korean	17	226,268
fa	Persian	17	213,903
tr	Turkish	17	149,718
hu	Hungarian	14	577,568
eo	Esperanto	14	105,286
hy	Armenian	10	127,988
cs	Czech	9	712,263
bg	Bulgarian	9	208,763
sq	Albanian	9	98,009
id	Indonesian	9	53,075
he	Hebrew	8	61,283
hr	Croatian	8	40,621
et	Estonian	7	20,405
sv	Swedish	6	270,642
no	Norwegian	6	62,767
az	Azerbaijani	5	52,762
fi	Finnish	4	60,507
tet	Tetum	4	18,485
lt	Lithuanian	3	16,572
mr	Marathi	3	16,386
hi	Hindi	3	3,490
ie	Interlingue	2	140,383
ta	Tamil	2	77,087
sw	Swahili	2	73,921
tl	Tagalog	2	35,962
gl	Galician	2	29,688
mk	Macedonian	2	14,654
th	Thai	1	70,909
tk	Turkmen	1	66,104
bs	Bosnian	1	63,018
kk	Kazakh	1	41,839
sl	Slovenian	1	22,844
sk	Slovak	1	12,997
co	Corsican	1	9,083
gn	Guarani	1	1,566
bo	Tibetan	1	579

Table 5: Language statistics including ISO-639 codes, number of documents, and number of mT5 tokens for the raw version of HALvest.

ISO-639	Language	# Documents	# mT5 Tokens
en	English	442,892	7,606,895,258
fr	French	193,437	8,728,722,255

Table 6: Language statistics including ISO-639 codes, number of documents, and number of mT5 tokens for the filtered version of HALvest.

Domain	Code	# Documents	# mT5 Tokens
Humanities and Social Sciences	shs	156,566	5,614,423,171
Computer Science	info	148,316	2,573,673,455
Life Sciences	sdv	115,744	3,145,323,780
Engineering Sciences	spi	102,751	2,254,653,825
Physics	phys	65,991	1,503,190,749
Mathematics	math	62,921	1,638,500,361
Chemical Science	chim	40,012	899,507,319
Environmental Science	sde	31,575	579,076,669
Sciences of the Universe	sdu	23,557	682,356,264
Cognitive Science	scco	11,772	227,487,096
Statistics	stat	10,579	184,678,350
Quantitative Finance	qfin	3,451	68,518,636
Nonlinear Sciences	nlin	1,972	30,694,088

Table 7: Domain statistics including domain codes, number of documents, and number of mT5 tokens for the raw version of HALvest.