Dynamic Reference Extraction and Linking across Multiple Scholarly Knowledge Graphs

Nicolau Duran-Silva[1,2], Pablo Accuosto[1],
[1]SIRIS Lab, Research Division of SIRIS Academic, Barcelona, Spain,
[2]LaSTUS Lab, TALN Group, Universitat Pompeu Fabra, Barcelona, Spain,

Abstract

References are an important feature of scientific literature; however, they are unstructured, heterogeneous, noisy, and often multilingual. We present a modular pipeline that leverages fine-tuned transformer models for reference location, classification, parsing, retrieval, and reranking across multiple scholarly knowledge graphs, with a focus on multilingual and non-traditional sources such as patents and policy documents. Our main contributions are: a unified pipeline for reference extraction and linking across diverse document types, openly released annotated datasets, fine-tuned models for each subtask, and evaluations across multiple scholarly knowledge graphs, enabling richer, more inclusive infrastructures for open research information.

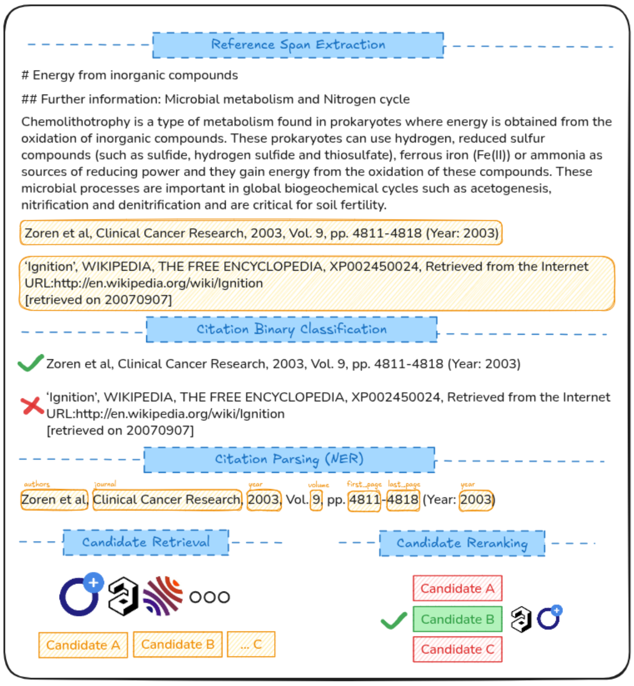

Figure 1: Overview of the pipeline and subtasks.

1 Introduction

Citations and references have been described as one of the most important features of scientific literature (Backes et al., 2024). They ground claims and reference previous work, connect research across disciplines, form the basis for the construction of scholarly knowledge graphs (SKGs), and enable bibliometrics and research impact evaluation and assessment (Leydesdorff et al., 2013; Cioffi and Peroni, 2022; Tkaczyk et al., 2018). Beyond scholarly articles, the number of documents that contain references to scientific work is increasing rapidly, ranging from project proposals, narrative CVs, patents, policy documents and public uses, and even social media and news (Lin et al., 2023; Cong et al.). In the context of open research information and open science, finding and linking references in multi-source documents is crucial for creating richer datasets and infrastructures.

However, extracting references from such diverse sources remains a challenge. Raw references appear in different citation styles (Tkaczyk et al., 2018), are often noisy or incomplete (missing DOI, title, or authors), and occur in multiple languages. Moreover, no single SKG offers complete coverage, making robust research object normalisation non-trivial.

Extraction and linking of scholarly references is an information extraction problem, and a key task of scholarly document processing (Backes et al., 2024). In an era of fake news and LLM hallucinations, research and new tools for grounding references and finding background support are fundamental. Existing tools focus mainly on parsing PDF articles. Although effective in controlled settings, they remain limited and are not very flexible in more diverse settings of references and document types. Recent experiments with LLMs (Backes et al., 2024) have shown mixed results, and previous research has underscored that deep-learning citation-parsing tools suffer from a lack of training data (Grennan and Beel, 2020).

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 80–86
December 23, 2025

In this work, we explore encoder-based language models for reference extraction and linking across multiple SKGs. We present a unified pipeline that combines reference location, reference parsing, retrieval, and re-ranking, and introduce ensemble-based linking to improve robustness across OpenAlex, OpenAIRE, CrossRef, and PubMed. To support this, we release new annotated datasets and fine-tuned models for each subtask, together with benchmark results demonstrating their effectiveness in multilingual and noisy-document settings. These resources enable reference extraction not only from scholarly articles but also from non-traditional sources, broadening the scope of SKG construction and downstream applications.

We have released our code, datasets and models fine-tuned in the context of this paper [1].

2 Related Work

A wide range of tools have been proposed for locating and parsing bibliographic references from PDF versions of scholarly articles (Cioffi and Peroni, 2022). Methods have relied on rule-based methods or shallow machine-learning approaches such as CRFs or SVMs (Zou et al., 2010; Tkaczyk et al., 2018), with widely used tools like ParsCit, AnyStyle, GROBID, CERMINE, Scholarcy, and Science Parse. Cioffi et al. (Cioffi and Peroni, 2022) differentiate between tools that can parse a single reference, those for parsing a list of references, and frameworks for parsing references from PDFs. Recent surveys (Backes et al., 2024; Cioffi and Peroni, 2022) report that GROBID and AnyStyle remain strong baselines, but also highlight that most tools focus on parsing rather than full extraction and linking, are restricted to a single database, and offer limited multilingual support. In addition, deep-learning approaches have been hindered by the lack of large annotated datasets (Grennan and Beel, 2020), and LLM-based attempts show mixed results (Backes et al., 2024). Biblio-Glutton (bib, 2018–2024) offers an open framework for reference resolution against authoritative records such as CrossRef, PubMed, HAL, and Unpaywall. While highly effective for processing scholarly articles, it remains tied to specific sources. In contrast, we explore encoder-based models designed to handle more diverse document types and reference settings.

[1] https://github.com/sirisacademic/references-tractor/

3 Materials and Methods

The modular pipeline comprises five steps (subtasks) to extract and link references, which are described below:

1. **Reference Location**: detect citation-bearing spans in raw documents (policy reports, patents, scholarly works, blogs), marking both the broader *citation-span* and the inline *citation-ref*, *author(s)*, *year*, and *citation-ID* (e.g., "(Smith et al., 2019)" or "[12]").

2. **Reference Classification**: the task of classifying citation-like text segments as academic references (e.g., journal articles, scholarly books, conference papers) or non-academic references (e.g., web pages, patents, generic abstracts). It is a binary classification that filters citations to scholarly works from other raw reference data, relevant for heterogeneous sources that cite a diverse set of documents.

3. **Reference Parsing (NER)**: a Named Entity Recognition (NER) model extracts key fields from the citation, parsing it into structured fields using a fine-tuned NER model. The extracted fields can include `TITLE`, `AUTHORS`, `VOLUME`, `ISSUE`, `YEAR`, `DOI`, `ISSN`, `ISBN`, `FIRST_PAGE`, `LAST_PAGE`, `JOURNAL`, and `EDITOR`.

4. **Reference Retrieval**: parsed fields are used to dynamically build queries to scholarly APIs.

5. **Reference Pairwise Reranking**: re-ranks pairs of the input reference and retrieved candidates from scholarly knowledge graphs.

3.1 Datasets

To support each component of the pipeline, we created five supervised datasets that cover the key subtasks: reference location, reference classification, reference parsing, pairwise reranking, and end-to-end multi-SKG linking. Table 1 provides an overview.

Dataset	Labels	Samples
Reference Location	5	1,922
Reference Classification	2	3,999
Reference Parsing (NER)	12	2,688
Reference Reranking	2	3,276
MultiSKG Linking	–	200

Table 1: Datasets overview.

Reference Location Dataset represents 1,922 annotated text segments from policy documents, patents, websites, news, and scientific papers, in both plain text and markdown formats. Each segment was manually annotated with the full citation span and the inline citation expression, enabling extraction of the reference span and its in-text context, including citation ID, year, and author mentions.

Reference Classification Dataset addresses the filtering step that separates scholarly citations from other sequences. We sample ∼5k non-patent literature entries from the PATSTAT database, covering common `NPL_TYPE` categories (`a`: unspecified, `b`: book, `s`: serial/journal, `w`: web). Each string is labeled TRUE (academic: journal article, scholarly book, conference paper, etc.) or FALSE (non-academic: web pages, office actions, manuals, etc.). Annotation follows a semi-supervised procedure: GPT-3.5 produces initial pseudo-labels, which we compare with the raw categories; we then split the corpus into two folds for cross pseudo-labelling, and human annotators resolve disagreements in Argilla (Daniel and Francisco, 2023) (see Appendix A, *Binary Classification prompt*). The final dataset is multilingual (mainly `en` and `zh`) and approximately balanced (55% TRUE, 45% FALSE), with a train/test split of 90/10.

Reference Parsing (NER) Dataset consists of 2,688 raw citation strings annotated with entity labels: `TITLE`, `AUTHORS`, `VOLUME`, `ISSUE`, `YEAR`, `DOI`, `ISSN`, `ISBN`, `FIRST_PAGE`, `LAST_PAGE`, `JOURNAL`, and `EDITOR`. The samples were gathered from non-patent literature entries in the PATSTAT database to ensure coverage of different citation formats and degrees of metadata completeness. The dataset is multilingual and was annotated following a semi-supervised approach. Pseudo-labels were generated with GPT-3.5 and refined by human annotators with Argilla (see Appendix A, *Reference Parsing (NER) prompt*).

Reference Pairwise Reranking Dataset provides 3,276 reference pairs. Each example is a pair of strings—raw reference and candidate—where the candidate is an APA-normalised reference constructed from OpenAlex metadata (authors, year, title, venue, volume, pages, DOI). Labels are binary (`1`=*same*; `0`=*different*). The corpus was built in two steps: (i) manual annotation of 1,276 candidate pairs to collect positive and hard negative examples, and (ii) to improve generalisation, we synthesise hard negatives by crossing citations with non-matching candidates.

MultiSKG Linking Dataset serves as a gold standard for end-to-end linking to multiple Scholarly Knowledge Graphs, we considered: OpenAlex (Priem et al., 2022), OpenAIRE (Manghi et al., 2012), CrossRef, and PubMed.[2] The dataset consists of 200 manually annotated references to the four target knowledge graphs providing unique identifiers for each source. Two annotators cross-annotated all references. Samples in the dataset vary in complexity, from well-structured to minimal metadata, including ambiguous and hard-to-match references, to evaluate real-world diversity.

3.2 Models & Training

We fine-tune transformer encoder models (Vaswani et al., 2017; Devlin et al., 2019), with model choices guided by baseline models (BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), DeBERTa-v3 (He et al., 2020), ModernBERT (Warner et al., 2025)), multilingual coverage (mBERT (Pires et al., 2019), XLM (Lample and Conneau, 2019)), and efficiency to support large-scale runs (multilingual DistilBERT (Sanh et al., 2019)). Our models are fine-tuned using the Hugging Face `Transformers` library, with early stopping and model selection based on validation performance. Hyperparameter configurations for each subtask (classification, NER, reranking) are reported in Appendix B.

3.3 Candidate Retrieval & Selection

The candidate retrieval component builds structured queries from the parsed citation fields and issues them to multiple scholarly knowledge graph APIs. Our approach includes:

- **Incremental metadata search:** Queries are constructed progressively, starting from high-confidence fields (e.g., DOI, title + year) and falling back to partial metadata combinations (e.g., authors + venue, title substrings) when primary identifiers are missing. We address this with multi-API retrieval, querying OpenAlex, OpenAIRE, Crossref, PubMed, and HAL, each offering different coverage, domain focus, and search capabilities.

- **Candidate reranking:** Retrieved candidates are scored with a fine-tuned pairwise model

[2]OpenAlex: `https://openalex.org/`, OpenAIRE: `https://explore.openaire.eu/`, Crossref: `https://www.crossref.org/`, and PubMed: `https://pubmed.ncbi.nlm.nih.gov/`.

(Section 3.1), which takes the raw reference and a candidate record as input and predicts whether they refer to the same publication. This learned approach combines lexical cues (title, authors, venue, year) with semantic similarity from transformer encoders, and the prediction score is used for reranking.

- **Ensemble linking:** After reranking, top-scoring candidates are cross-compared across APIs. When DOIs are present, we perform a majority-vote consensus to mitigate single-API inconsistencies and maximise coverage.

4 Evaluation

4.1 Experimental Setup

Our datasets, described in Section 3.1, were split 80/10/10 into train, development, and test. Models were fine-tuned as described in Section 3.2. We report results using macro-F1, computed on the held-out test split. For NER tasks, we compute token-level F1 scores on entity spans. For reference linking, we evaluate on the MultiSKG dataset by requiring exact DOI/ID matches as correct.

4.1.1 Task-level Evaluation

Model	Location	Classification	Parsing	Reranking
DistilBERTm	.755	.935	.949	.904
BERTm-base	.773	**.944**	.957	.902
RoBERTa-base	.788	.940	**.962**	**.915**
XLM-base	–	.914	.957	.901
DeBERTa-v3-base	**.792**	.932	.961	.903
ModernBERT	.732	.936	.955	**.915**

Table 2: Task-level results across models (macro-F1).

We first evaluate each subtask independently. Table 2 shows that RoBERTa and DeBERTa-v3 achieve consistently strong performance across NER and reranking, while BERTm provides the best overall performance on the classification task. DistilBERT offers competitive results with lower computational cost, and XLM demonstrates robust multilingual generalisation.

4.2 Linking Evaluation

We evaluate per-API accuracy with an error breakdown. As shown in the results in Table 3, the ensemble achieves the highest accuracy.

5 Discussion

While overall performance across the subtasks is strong, the linking evaluation reveals several ambiguous cases that complicate strict accuracy metrics. Many of the errors occur during the reranking step and are actually ambiguous matches: although correct DOIs are often retrieved, metadata mismatches (e.g., page ranges, abbreviated venues, missing affiliations) can lead to false negatives. For example, "*Yamagishi et al., J. Phycol. 43: 519–527 (2007)*" illustrates how strict page-number matching can cause the reranker to fail, even when the DOI is correct. Additional errors arise from different versions or duplicate entries with different unique IDs, suggesting that recall-based evaluation might better reflect the system's performance. Some errors are due to partial parsing, while others are caused by missing records in certain SKGs. While the pipeline's true impact lies in its ability to handle cross-database complexities, improving the reranking step would result in better handling of ambiguous matches.

API	Accuracy	C_Match	I_Miss	I_Match
OpenAlex	.745	127	15	30
OpenAIRE	.675	105	19	34
PubMed	.590	48	12	5
CrossRef	.640	104	23	39
Ensemble	**.755**	122	24	19

Table 3: Linking evaluation results, reporting accuracy on strict DOI/ID match. Error breakdown as C_Match (correct matches), C_NoRes (correct empty), I_Miss (missed matches), and I_Match (incorrect matches).

6 Conclusions

We propose a novel pipeline for multilingual reference extraction and linking, using fine-tuned transformer models to enhance scholarly knowledge graph coverage. The approach combines transformer models, incremental retrieval, and ensemble reranking for robust performance in noisy, multilingual settings. We aim to create open citation datasets from policy documents and patents, and expand linking to national and discipline-specific SKGs. Future work will focus on scaling for larger datasets and exploring span-based techniques and long-context models to improve citation extraction from lengthy documents, broadening its applicability to open research infrastructures.

7 Limitations and Future Work

While the proposed pipeline demonstrates strong performance across individual subtasks, several limitations guide ongoing development. The datasets are relatively small (1,922 spans for location, 2,688 for NER, 3,276 for reranking, and 200 gold references for multi-KG linking) and rely on semi-supervised annotation with GPT-based pseudolabels and human adjudication. Larger, more diverse datasets with reported inter-annotator agreement are needed to strengthen claims of generalization across domains and languages.

Our end-to-end evaluation uses strict DOI/ID matching on a limited multilingual sample. As discussed in Section 5, many errors arise from metadata inconsistencies across knowledge graphs rather than true matching failures. Future work should incorporate relaxed matching criteria, additional metrics (top-k recall, MRR), and systematic comparison with established reference extraction systems on standardized benchmarks.

The pipeline assumes text or markdown input and does not explicitly handle PDF layout or OCR errors, which limits applicability to certain document types. Integration with PDF extraction tools would broaden the scope. Additionally, the reranking component could be improved to better handle metadata ambiguity (abbreviated venues, page range variations) through fuzzy matching and multi-field attention. Finally, explicit mechanisms for detecting potentially fabricated or hallucinated references would strengthen the system's reliability.

All models, code, and datasets are openly available, and ongoing experiments will be progressively added to the project repository.

Acknowledgments

Supported by the Industrial Doctorates Plan of the Department of Research and Universities of the Generalitat de Catalunya, by Departament de Recerca i Universitats de la Generalitat de Catalunya (ajuts SGR-Cat 2021). This work was co-funded by the EU HORIZON SciLake (Grant Agreement 101058573). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the funders.

References

2018–2024. biblio-glutton. *Preprint*, swh:1:dir:a5a4585625424d7c7428654dbe863837aeda8fa7.

Tobias Backes, Anastasiia Iurshina, Muhammad Ahsan Shahid, and Philipp Mayr. 2024. Comparing free reference extraction pipelines. *International Journal on Digital Libraries*, 25(4):841–853.

Alessia Cioffi and Silvio Peroni. 2022. Structured references from pdf articles: assessing the tools for bibliographic reference extraction and parsing. In *International Conference on Theory and Practice of Digital Libraries*, pages 425–432. Springer.

Ting Cong, Er-Te Zheng, Zekun Han, Zhichao Fang, and Rodrigo Costas. Social media uptake of scientific journals: A comparison between x and wechat. *Journal of Information Science*, page 01655515251359759.

Vila-Suero Daniel and Aranda Francisco. 2023. Argilla - Open-source framework for data-centric NLP.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Mark Grennan and Joeran Beel. 2020. Synthetic vs. real reference strings for citation parsing, and the importance of re-training and out-of-sample data for meaningful evaluations: Experiments with GROBID, GIANT and CORA. In *Proceedings of the 8th International Workshop on Mining Scientific Publications*, pages 27–35, Wuhan, China. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Loet Leydesdorff, Ismael Rafols, and Chaomei Chen. 2013. Interactive overlays of journals and the measurement of interdisciplinarity on the basis of aggregated journal–journal citations. *Journal of the American society for Information science and Technology*, 64(12):2573–2586.

Zihang Lin, Yian Yin, Lu Liu, and Dashun Wang. 2023. Sciscinet: A large-scale open data lake for the science of science research. *Scientific Data*, 10(1):315.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Paolo Manghi, Nikos Houssos, Marko Mikulicic, and Brigitte Jörg. 2012. The data model of the openaire scientific communication e-infrastructure. In *Research Conference on Metadata and Semantic Research*, pages 168–180. Springer.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

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*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Dominika Tkaczyk, Andrew Collins, Paraic Sheridan, and Joeran Beel. 2018. Evaluation and comparison of open source bibliographic reference parsers: A business use case.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Jie Zou, Daniel Le, and George R Thoma. 2010. Locating and parsing bibliographic references in HTML medical articles. *International Journal on Document Analysis and Recognition (IJDAR)*, 13(2):107–119.

A Pseudo-Annotation Prompts

Binary Classification prompt

```
Given a piece of text, classify it into one of the
      following categories:

- TRUE: include publications, review, "academic" book
       chapters, conference papers.
- FALSE: references to other types of sources

Instructions: Analyze the content of the provided
      text and assign the appropriate category,
      returning TRUE or FALSE

Examples of TRUE:
- Matthews et al., Homeostasis Model Assessment:
      Insulin Resistance and Beta-cell Function from
      Fasting Plasma Glucose and Insulin
      Concentrations in Man, Diabetologia, 28, (1985),
       pp. 412-419.
- Liu, et al.; Design of carbonylative polymerization
       of heterocycles. Synthesis of polyesters and
      poly(amide-block-ester)s J. Am. Chem. Soc. 2004,
       vol. 126, pp. 14716-14717; 6 pages.
- JULIAN FIERREZ-AGUILAR ET AL: 'Incorporating Image
      Quality in Multi-algorithm Fingerprint
      Verification', 1 January 2005, ADVANCES IN
      BIOMETRICS LECTURE NOTES IN COMPUTER SCIENCE;;
      LNCS, SPRINGER, BERLIN, DE, PAGE(S) 213 - 220,
      ISBN: 978-3-540-31111-9, XP019026878
- KUANG-HUA CHANG: 'E-Design computer-Aided
      Engineering Design', 2015, ELSEVIER ACADEMIC
      PRESS
- HWANG J S ET AL: 'Heteroepitaxy of gallium nitride
      on (0001), (1012) and (1010) sapphire surfaces',
       JOURNAL OF CRYSTAL GROWTH, ELSEVIER, AMSTERDAM,
       NL LNKD- DOI:10.1016/0022-0248(94)90263-1, vol.
       142, no. 1-2, 1 September 1994 (1994-09-01),
      pages 5 - 14, XP024439721, ISSN: 0022-0248, [
      retrieved on 19940901]
- KUMA HIROYUKI ET AL: 'Liquid phase immunoassays
      utilizing magnetic markers and SQUID
      magnetometer', CLINICAL CHEMISTRY AND
      LABORATORY MEDICINE, vol. 48, no. 9, 1 January
      2010 (2010-01-01), DE, XP055783197, ISSN:
      1434-6621, Retrieved from the Internet <URL:
      http://dx.doi.org/10.1515/CCLM.2010.259> DOI:
      10.1515/CCLM.2010.259
- PAN ET AL.: 'Sustainable production of highly
      conductive multilayer graphene ink for wireless
       connectivity and loT applications', NATURE
      COMMS, vol. 9, 2018, pages 5197
- LELOIR, L.F., ARCH BIOCHEM, vol. 33, no. 2, 1951,
      pages 186 - 90

Examples of FALSE:
- Final Office Action, U.S. Appl. No. 13/316,351,
      dated Jul. 31, 2013, 20 pages.
- U.S. Appl. No. 13/006,270, filed Jan. 13, 2011 Non-
      Final Office Action dated Sep. 12, 2014, 41
      pages.
- Matrx Metalloproteinase, from Wikipedia,the free
      encyclopedia (8 pages), retrieved from the
      Internet on Dec. 17, 2009 at http://en.
      wikipedia.org/wiki/Matrix-metalloproteinase.
- 'Double Layer DVD+R Multi-Media Command Set
      Description, Version 1.00', 4 June 2004, ROYAL
      PHILIPS ELECTRONICS, EINDHOVEN, THE NETHERLANDS,
       XP002386267
- 'The Leukocyte Antigen Facts Book', 1997, HARCOURT
      BRACE & CO.
- DOUGLAS GRAHAM: 'Folding a bandana into fade mask',
       6 April 2020 (2020-04-06), XP055859991,
      Retrieved from the Internet <URL:https://www.
      youtube.com/watch?v=dI3343Gb9YA> [retrieved on
      20211110]
- Banknote Paper', WEBPAGES G&D, pages 9PP,
      XP055351061, Retrieved from the Internet <URL:
      https://www.gi-de.com/en/products_and_solutions/
      products/banknote_paper/banknote-paper.jsp>
- PHILIPS: 'Fallback mode for Rel-7 FDD MIMO scheme',
       3GPP TSG RAN WG1 MEETING #46 TDOC R1-061952

Predict the category for this text:
{INPUT_TEXT}
```

Reference Parsing (NER) prompt

```
Can you parse this citation string:
"{INPUT_TEXT}"

in the following attributes:
- authors
- title
- editor
- volume
- issue
- publication date
- publisher
- journal
- first_page
- last_page
- doi
- isbn
- issn
- link online

Only return attributes in bullet points with a not
      empty value
```

B Fine-tuning hyperparameters

B.1 Text Classification

We fine-tune transformer encoder models for the **Reference Classification** task by adding a classification head with two output labels, implemented with HuggingFace `Transformers`. Each model was trained on a single NVIDIA A100 GPU for up to 6 epochs with early stopping (patience 2) with main hyperparameters described in Table 4.

Hyper-parameter	**Value**
Learning Rate	2e-5
Learning Rate Decay	Linear
Weight Decay	0.01
Warmup Steps	0
Batch Size	32
Max. Training Epochs	6
Metric for best model	F1-macro

Table 4: Fine-tuning hyperparameters for the Reference Classification task.

B.2 NER

For the **Reference Location** and **Reference Parsing** tasks, we fine-tune transformer encoder models with a token classification head, using subword-level alignment. All models were trained on a single NVIDIA A100 GPU with early stopping (patience 2). Table 5 summarises the main hyperparameters.

Hyper-parameter	**Value**
Learning Rate	2e-5
Learning Rate Decay	Linear
Weight Decay	0.01
Warmup Steps	0
Batch Size	32
Max. Training Epochs	25
Max Sequence Length	512
Metric for best model	F1
Early Stopping Patience	2

Table 5: Fine-tuning hyperparameters for the Reference Location and Reference Parsing tasks.

B.3 Pairwise Reranking

For the **Pairwise Reranking** task, the goal is to classify pairs of references (`reference1`, `reference2`) as either referring to the same publication (`1`) or different publications (`0`). Each pair is encoded as a single sequence by concatenating the two reference strings with a special separator token (`[SEP]`). We fine-tune transformer encoder models with a sequence classification head (two output labels). Models were trained on a single NVIDIA A100 GPU with early stopping (patience 2). Table 4 reports the training hyperparameters.