An Enhanced Training-Free Pipeline for Entity Recognition and Linking: A Low-Resource Case Study – 20-th Century Historical Medical Texts

Phu-Vinh Nguyen
Uppsala University, Sweden
phu-vinh.nguyen.1216@student.uu.se

Vera Danilova
Uppsala University, Sweden
vera.danilova@idehist.uu.se

Abstract

Entity linking in biomedicine typically relies on large annotated corpora and supervised methods, which often fail in out-of-distribution settings. Historical medical texts are rich in biomedical terms but pose unique challenges: terminology has changed, some concepts are obsolete, and stylistic differences from modern journals prevent off-the-shelf models fine-tuned on contemporary datasets from aligning historical terms with current ontologies. Training-free methods based on LLMs offer a solution by linking historical terms to modern concepts and inferring their meaning from context. In this paper, we evaluate a state-of-the-art training-free entity linking method on historical medical texts and propose an improved pipeline—end-to-end entity extraction and linking with confidence estimation. We also assess performance on modern benchmarks to check whether the gains generalize to other domains and show their superior performance in most cases. We report an analysis of the findings. The code and curated dataset for historical medical entity linking are available on GitHub[1]

1 Introduction

Historical medical texts preserve medical knowledge, offering insights for both historians and medical professionals. They document observations and therapies relevant to ethnopharmacological and comparative biomedical research (Connelly et al., 2020) and inform broader medical practice (Patel and Desai, 2014; Hays, 2024). Integrating such knowledge with modern evidence requires establishing semantic links between historical terminologies and contemporary medical concepts, a task performed in NLP by Entity Linking (EL).

Modern biomedical EL pipelines, based on bi-encoders, cross-encoders, or reinforcement learning (Gillick et al., 2019; Gupta et al., 2017; Sevgili et al., 2020; Broscheit, 2019; Kolitsas et al., 2018; Agarwal and Bikel, 2020) and trained on contemporary corpora like MedMentions (Mohan and Li, 2019) or COMETA (Basaldella et al., 2020), struggle in out-of-domain (OOD), resource-scarce settings. Historical texts exacerbate this problem due to temporal shifts in lexical form and meaning, as well as distinctive stylistic conventions.

Earlier entity linking efforts for historical medical texts (Thompson et al., 2016) rely on custom schemas and manually curated term inventories, which are labor-intensive, ad-hoc, and language-dependent. Large language models (LLMs) have shown promise in overcoming these limitations. For example, Fillies et al. (2025) demonstrated that LLMs can address key challenges in historical species naming, including spelling changes, new terms, shifts between broad and specific names, and renaming of common names. This makes them more promising for new and low-resource tasks, where large training corpora are not feasible.

Motivated by these findings, we investigate LLM-based entity linking for historical medical texts by evaluating a state-of-the-art few-shot EL framework OneNet (Liu et al., 2024), and proposing an alternative method augmented with automatic candidate extraction and confidence-aware linking that outperforms OneNet on our newly constructed dataset, as well as across several common benchmarks.

Our key contributions are:

1. *A new enhanced training-free EL pipeline for low-resource settings.*
2. *Evaluation of the new pipeline on the dataset and extensive comparison to the state-of-the-art baseline.*
3. *A curated dataset of historical medical texts for entity linking.*

[1] https://github.com/ActDisease/CbEL—Entity-Linking-for-Historical-Medical-Texts

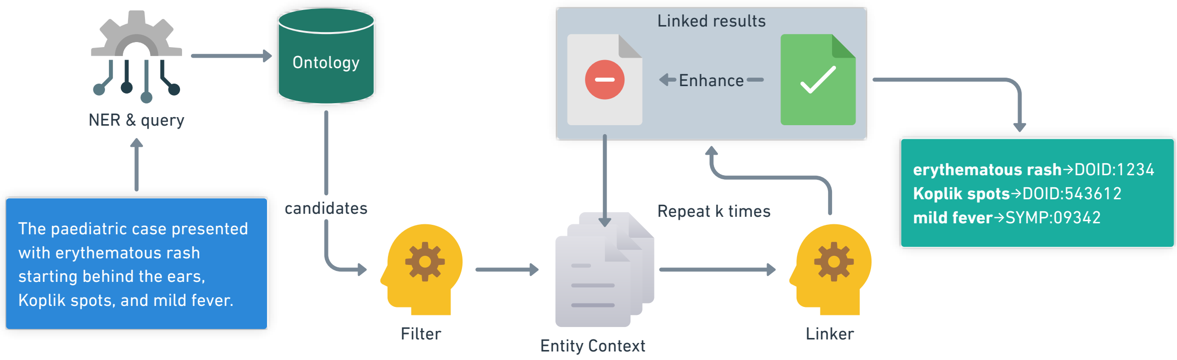

Figure 1: Our entity linking pipeline that considers LLMs' confidence and verified entities in the linking process.

2 Related work

Information extraction from historical medical periodicals was first systematically addressed by Thompson et al. (2016), who developed dedicated semantic resources for medical history text mining. Their work introduced the HIMERA corpus, consisting of expert annotations of 35 British Medical Journal articles using a custom schema of broad, historian-defined entity types, and a time-sensitive terminological resource derived from two British medical archives. Historical terminology was first linked to UMLS (Bodenreider, 2004) through exact and fuzzy matching. Over 60% of historically relevant variants of UMLS concepts could not be automatically aligned. Next, distributional semantic modeling was used to capture diachronic variation. While effective, this approach depends on substantial corpora and expert effort, and is difficult to scale across time periods, domains, or ontologies.

Existing EL solutions for the biomedical domain include MedLinker (Loureiro and Jorge, 2020) or BioBART (Yuan et al., 2022). However, their limitation is that they require a large corpus for fine-tuning, which is not suitable for low-resource domains. More recently, LLMs have emerged as a data-efficient alternative for text processing. Previous work shows that zero-shot and few-shot prompting can achieve competitive entity recognition and linking performance on historical corpora, effectively handling key challenges of historical texts, including spelling variation, stylistic divergence, and semantic shift (Zhang and Colavizza, 2025; Boscariol et al., 2025; Fillies et al., 2025). This highlights their potential for EL through context-aware reasoning in the historical medical domain and motivates further analysis of this approach.

3 Baseline Approach

OneNet (Liu et al., 2024) is a state-of-the-art few-shot EL pipeline designed to address low-resource and OOD scenarios that has been validated on several major modern EL benchmarks, such as AIDA-CoNLL (Hoffart et al., 2011) and MSNBC (Cucerzan, 2007). It combines candidate filtering, in-context and standalone linking, and final answer verification, which employs LLM to resolve disagreement between the two linking styles. OneNet's full algorithm is shown in Algorithm 2.

Despite its strong performance, the framework has several limitations relevant to our case study. First, it assumes that entity mentions are given, lacks a named entity recognition component, and thus prevents fully automatic end-to-end processing. Second, it does not account for the LLM's confidence, enforcing a link even under high uncertainty. Those limitations hinder the transparency and reliability of EL results. In our work, we combine this baseline with NER models to evaluate on the full entity linking problem, not just entity disambiguation.

4 Our Method

To address the limitations of the baseline, we propose a pipeline called *CbEL*, short for confidence-based entity linking. This pipeline comprises three main stages: entity detection and short description generation, candidate search, and disambiguation using LLM confidence scores (Fig. 1). In the first stage, a Named Entity Recognition (NER) method is used to detect entities in the text, and then LLMs generate a list of candidate keywords for searching each entity in a knowledge base (KB). After generating the keyword list, we retrieve similar candidate terms from the KB with methods like fuzzy

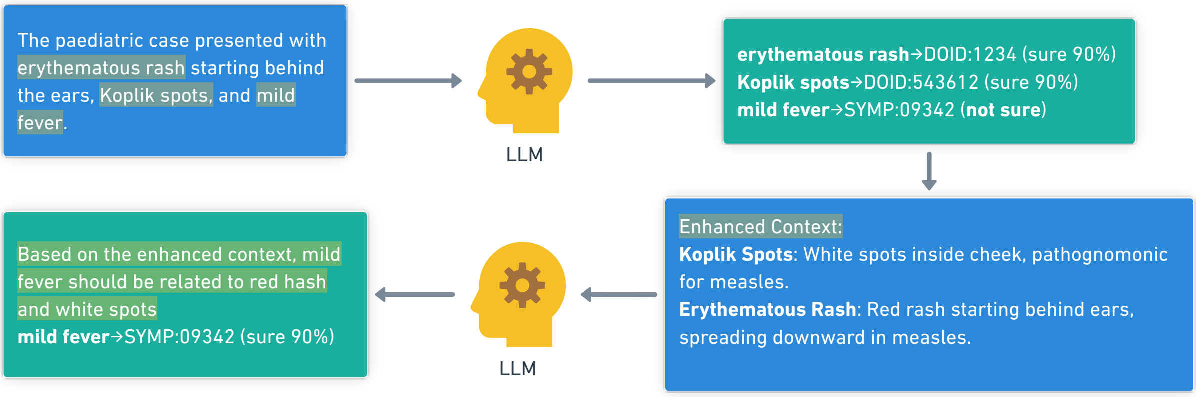

Figure 2: An example of how enhanced context is used for ambiguous and low-confidence entities.

matching and n-gram search. These terms serve as candidates for the subsequent confidence-based EL step, which will be filtered using a similar process to OneNet. Here, the LLM is **prompted** to select the top-k candidates most relevant to each detected entity and to provide a confidence score for each candidate, reflecting its belief that the candidate is correct. However, only the most confident candidate is selected as the linked result. Given a list of high-confidence entities after linking, we return these results and utilize their information to enrich the context for linking low-confidence entities, as shown in Fig. 2. By iterating this process, we mitigate ambiguity for difficult entities by leveraging the definitions of already-linked ones. Ultimately, only high-confidence entities are returned, and uncertain or incorrect entities are erased from the final result. The process of the pipeline is summarized in Algorithm 1, where $conf$ denotes the confidence extraction process by prompting LLM to generate an answer with its confidence in the answer from 0 to 1. In our work, the threshold τ is set to 0.75.

Algorithm 1 Confidence-based Entity Linking

Require: Document D, threshold τ, iterations K
Ensure: High-confidence entity links $\mathcal{L}$

```
ℰ ← ExtractEntitiesAndCandidates(D)
ℒ ← ∅
𝒞 ← D {Initial context}
for t = 1 to K do
    ℋ ← {(e, id) | e ∈ ℰ, conf(e, id, 𝒞) ≥ τ}
    ℒ ← ℒ ∪ ℋ
    𝒞 ← 𝒞 ∪ ℋ {Enrich with (entity, id) pairs}
    ℰ ← ℰ \ {e | (e, id) ∈ ℋ}
end for
return ℒ
```

5 MedHistory Dataset

The dataset [2] for annotation currently comprises 514 texts from the *British Medical Journal* (BMJ), sampled from publications across different time periods in the 20th century, with full text extracted using DeepSeek-OCR (Wei et al., 2025). Annotation is ongoing voluntarily, involving nine experts (seven medical doctors and two historians of medicine). A dedicated web interface (Fig. 3 in Appendix) was developed to support entity linking to the Human Disease Ontology (DO) (Schriml et al., 2025), enhanced with key functionalities of semantic similarity-based concept search and historical term flagging, as well as additional functions requested by the annotators, such as comments, typo reporting, and access to original pages.

The DO is an open-source, hierarchically structured resource covering over 10k human diseases, organized by etiology and enriched with cross-ontology mappings and obsolete terminology. Its hierarchical structure facilitates studying how disease concepts split, merge, or are redefined over time. Its core classes closely align with the historically motivated schema proposed by Thompson et al. (2016). This makes it well-suited for aligning historically variable disease terminology.

A detailed guideline was developed for annotators, providing step-by-step instructions on using the web-based interface, searching and selecting ontology entities, and applying annotation flags. The guide also specifies annotation conventions, span selection criteria, disambiguation principles, handling of historical or ambiguous terms, and examples of common edge cases to ensure consistency across annotators.

[2] https://huggingface.co/datasets/npvinHnivqn/MedHistory

The tool allows shared visibility of labeled samples, without revealing label names, and supports labeling of the same or overlapping spans. To date, 235 unique entities have been manually annotated, forming the basis for our evaluation.

6 Experiments

Our experiments pursue two objectives: (1) evaluating the baseline and our method on MedHistory, including a subset explicitly flagged as historical, and (2) assessing generalization on contemporary benchmarks, namely the NCBI Disease Corpus[3] and standard news-domain EL datasets annotated with Wikipedia entities (AIDA-CoNLL, MSNBC, KORE50 (Hoffart et al., 2012)). This setup allows us to assess the LLM's ability to handle contemporary disease terminology and to examine whether the observed performance gains generalize across domains. The statistics of all benchmarks being used in this experiment are reported in Table 2.

Due to the nature of the entity linking problem, which includes NER and entity disambiguation, we report results on each task separately to enable direct comparison with OneNet on different aspects using the standard micro F1 metric, and additionally evaluate end-to-end pipeline performance. The evaluation metrics are computed after the pipeline finishes linking entities in a document, allowing the assessment of the entire pipeline and combination, not just evaluating modules separately. All experiments use the Qwen-3-8B model (Team, 2025) with 4-bit quantization and the same pretrained NER model for both methods to ensure a fair comparison. Lastly, we compare only against the state-of-the-art training-free method, excluding fine-tuned approaches.

7 Results

Overall, the results depicted in Table 1 reveal complementary strengths across datasets and tasks. Because our method (CbEL) applies confidence-based filtering, incorrectly detected entities from the NER model are removed in our pipeline, reducing the false positive rate compared to OneNet. For example, while the NER model might detect 'Parkinson' as a disease, the LLM uses contextual understanding to flag this term as a person's name and remove it from the detected entities, thereby enhancing both NER accuracy and overall pipeline performance. Owing to its improved NER performance, our method outperforms OneNet in overall results by reducing the number of unlinkable entities. NER performance is consistently higher on modern benchmarks but degrades substantially on historical entities, reflecting the increased difficulty posed by lexical and conceptual drift. We report results on both the full historical dataset (MedHistory) and on the subset of entities explicitly flagged as historical (MedHistory-hs), which captures terms that are barely mentioned in modern documents and mostly used in the 20th century.

For EL (highlighted in bold), the highest micro F1 scores are observed on the modern biomedical corpus (NCBI) and on the MedHistory-hs subset (54 entities). In the biomedical setting, OneNet and CbEL achieve comparable performance, with CbEL showing a modest advantage on historical medical texts. Notably, CbEL substantially outperforms OneNet on news-domain benchmarks, indicating that the proposed solution generalizes effectively beyond the medical domain.

Examining failure cases for both methods, we find that, on average, CbEL generates 90 candidate entities, of which approximately 12% are correctly linked, with only 3 incorrect assignments. OneNet produces on average 290 candidates to recover just 3 additional correct entities while introducing 5 errors, resulting in substantially lower overall metrics. Both approaches consistently fail to recognize historical terms such as "increased richness of the blood" or "Bright's disease", highlighting the challenges posed by terminology unfamiliar to LLMs and NER models and underscoring the importance of tailored methods for historical medical texts.

8 Conclusion

This paper proposes a novel EL approach that integrates uncertainty quantification with known entities. Through experiments on historical medical texts and modern benchmarks, we demonstrate that CbEL outperforms OneNet across various linking tasks, including the low-resource historical setting. We also present a carefully annotated benchmark for this domain, providing a valuable resource for future evaluation. While our pipeline leverages a general-purpose LLM, its limited coverage of historical medical terminology constrains performance. Furthermore, current solutions cannot resolve a large number of candidates. Those problems will need to be addressed in future work.

[3] https://www.ncbi.nlm.nih.gov/research/bionlp/Data/disease/

Table 1: **Evaluation results.** MedHistory-hs is a sub-dataset of MedHistory in which terms are flagged as historical, which barely exist in the modern documents. **Bold values** indicate the higher disambiguation results of the two methods.

Benchmark	Type	CbEL			OneNet		
		NER F1	Disamb F1	Full F1	NER F1	Disamb F1	Full F1
MedHistory	Medical	0.294	**0.500**	0.140	0.192	0.492	0.123
MedHistory-hs	Medical	0.194	**0.785**	0.152	0.114	0.736	0.084
NCBI	Medical	0.668	0.778	0.422	0.537	**0.780**	0.437
aida-conll	Common	0.646	**0.510**	0.330	0.664	0.292	0.194
kore50	Common	0.829	**0.526**	0.437	0.834	0.181	0.151
msnbc	Common	0.659	**0.688**	0.453	0.652	0.406	0.264

Limitations

While the pipeline demonstrates improved performance by leveraging confidence and enhanced context from linked entities, this work has some limitations. First, the heart of the pipeline is LLM, whose performance fluctuates with different prompting techniques and LLMs. However, this limitation is known to all LLM-based pipelines. Secondly, despite the pipeline providing a confidence score to explain its prediction, LLMs maintain a black box, which cannot be transparent, just like other deep learning solutions.

Potential Risks

The pipeline utilizes LLMs and deep-learning methods for entity linking, which still suffer from hallucinations. Consequently, the pipeline should only be used as an assisting tool.

Acknowledgments

This research was supported by the ActDisease project at Uppsala University. Computational resources were provided by the National Infrastructure for Supercomputing in Sweden (NAISS). We are grateful to Henrik Klasson, Anders Backlund, Andrew Burchell, Diego Alonso Ortiz Ortigas, Klara Lithammer, Tryggve Nevéus, and Ylva Söderfeldt for their valuable contributions to dataset preparation.

References

Oshin Agarwal and Daniel M. Bikel. 2020. Entity linking via dual and cross-attention encoders. *ArXiv*, abs/2004.03555.

Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. COMETA: A corpus for medical entity linking in the social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3122–3137, Online. Association for Computational Linguistics.

Hannah Bast, Matthias Hertel, and Natalie Prange. 2022. ELEVANT: A fully automatic fine-grained entity linking evaluation and analysis tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 72–79, Abu Dhabi, UAE. Association for Computational Linguistics.

O. Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–D270.

Marta Boscariol, Luana Bulla, Lia Draetta, Beatrice Fiumanò, Emanuele Lenzi, and Leonardo Piano. 2025. Evaluation of llms on long-tail entity linking in historical documents. *Preprint*, arXiv:2505.03473.

Samuel Broscheit. 2019. Investigating entity knowledge in BERT with simple neural end-to-end entity linking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 677–685, Hong Kong, China. Association for Computational Linguistics.

Erin Connelly, Charo I. del Genio, and Freya Harrison. 2020. Data mining a medieval medical text reveals patterns in ingredient choice that reflect biological activity against infectious agents. *mBio*, 11(1):10.1128/mbio.03136–19.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic. Association for Computational Linguistics.

Jan Fillies, Maximilian Teich, Naouel Karam, Adrian Paschke, and Malte Rehbein. 2025. Historic to fair: Leveraging llms for historic term identification and standardizationhistorisch zu fair: Einsatz von llms zur identifikation und standardisierung historischer begriffe. *Datenbank-Spektrum*.

Daniel Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldridge, Eugene Ie, and Diego Garcia-Olano. 2019. Learning dense representations for entity retrieval. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 528–537, Hong Kong, China. Association for Computational Linguistics.

Nitish Gupta, Sameer Singh, and Dan Roth. 2017. Entity linking via joint encoding of types, descriptions, and context. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2681–2690, Copenhagen, Denmark. Association for Computational Linguistics.

Richards Hays. 2024. The relevance of medical history to current practice. *Australian Journal of General Practice*, 53(3):157–160.

Johannes Hoffart, Stephan Seufert, Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. 2012. KORE: keyphrase overlap relatedness for entity disambiguation. In *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, pages 545–554. ACM.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-end neural entity linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529, Brussels, Belgium. Association for Computational Linguistics.

Xukai Liu, Ye Liu, Kai Zhang, Kehang Wang, Qi Liu, and Enhong Chen. 2024. OneNet: A fine-tuning free framework for few-shot entity linking via large language model prompting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13634–13651, Miami, Florida, USA. Association for Computational Linguistics.

Daniel Loureiro and Alípio Mário Jorge. 2020. Medlinker: Medical entity linking with neural representations and dictionary matching. In *Advances in Information Retrieval*, volume 12036 of *Lecture Notes in Computer Science*, pages 230–237, Cham. Springer International Publishing.

Sunil Mohan and Donghui Li. 2019. Medmentions: A large biomedical corpus annotated with umls concepts. *ArXiv*, abs/1902.09476.

Parth M. Patel and Sukumar P. Desai. 2014. A clinician's rationale for the study of history of medicine. *Journal of Education in Perioperative Medicine*, 16(4):E070.

Lynn M. Schriml, J. Allen Baron, Claudia Marie Sánchez-Beato Johnson, James B. Munro, Elvira, Sue Bello, Melody Swen, Anu, Becky Jackson, Andra Waag, Chris Mungall, and James A. Overton. 2025. Diseaseontology/humandiseaseontology: Do november 2025 release (v2025-11-25). OBO Foundry ontology, DOID.

Ozge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Christian Biemann. 2020. Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 13:527–570.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Paul Thompson, Riza Theresa Batista-Navarro, Georgios Kontonatsios, Jacob Carter, Elizabeth Toon, John McNaught, Carsten Timmermann, Michael Worboys, and Sophia Ananiadou. 2016. Text mining the history of medicine. *PLOS ONE*, 11(1):e0144717.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression. *ArXiv*, abs/2510.18234.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. BioBART: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109, Dublin, Ireland. Association for Computational Linguistics.

Shibingfeng Zhang and Giovanni Colavizza. 2025. Named entity recognition of historical text via large language model. *Preprint*, arXiv:2508.18090.

A Baseline

This section provides the algorithm of the training-free baseline that we compare our method with. Overall, the method will first use LLMs to summarize and filter irrelevant candidates using the ERP (Entity Reduction Processor). After that, the method will link entities using contextual knowledge (input paragraph) and prior knowledge (LLMs' knowledge). If both predictions are not the same, a resolver (LLM) will be used with prior knowledge and context to get the final entity.

Algorithm 2 OneNet: Fine-Tuning Free Entity Linking Framework

Require: Mention m, Context S, Candidate Entities $\theta = \{e_1, \ldots, e_n\}$
Ensure: Linked Entity e^*

{— **Entity Reduction Processor** —}
$\theta_{sum} \leftarrow$ **SummarizeEntityDescriptions**(θ)
$\theta_{filtered} \leftarrow$ **FilterIrrelevantEntities**(m, S, θ_{sum})
{— **Dual-Perspective Entity Linker** —}
$e_{context} \leftarrow$ **ContextualEntityLinking**$(m, S, \theta_{filtered})$
$e_{prior} \leftarrow$ **PriorKnowledgeLinking**$(m, \theta_{filtered})$
{— **Entity Consensus Judger** —}
if $e_{context}$ matches e_{prior} **then**
    $e^* \leftarrow e_{context}$
else
    $e^* \leftarrow$ **ResolveDisagreement**$(m, S, e_{context}, e_{prior})$
end if
return e^*

B Settings

In this paper, we use Qwen-3.1-8B (Team, 2025) as a standard generative language model for all experiments, ensuring a fair evaluation of each pipeline. Furthermore, this choice is based on the balance between performance and the resource efficiency of this model. This LLM choice might lead to lower performance of the OneNet, but it provides better pipeline comparison in general. Next, for CbEL, we limit the repeating time k to 3, which helps accelerate the algorithm.

Furthermore, in the experiment, we report three main Micro F1 scores, including NER, which is the performance of the NER model, disambiguation, which is the performance of both our entity recommendation method (fuzzy matching) and the entity disambiguation module of each pipeline, and the F1 score of the full pipeline. This is the reason why the performance is very low compared to the entity disambiguation problem alone, where candidates, including the correct candidate, and a list of correctly detected entities are provided. Lastly, we used the ELEVANT (Bast et al., 2022) repository to support our experiments.

To solve the NER problem, we used three models, including *en_core_web_lg* for the general domain, *en_ner_bc5cdr_md* for diseases, and *en_ner_jnlpba_md* for genes. This is applied to both OneNet and our pipeline to ensure fair judgment.

C Dataset Description

AIDA-CoNLL is a news-based entity linking benchmark with 231 articles from the 1990s, manually annotated with YAGO2 entities. Its specialization lies in short, easily-detectable mentions (94.5% are 1-2 words) concentrated in sports content (44% of articles), creating a domain-biased but widely-used Wikipedia evaluation standard.

KORE50 comprises 50 handcrafted sentences emphasizing challenging disambiguation across five domains (celebrities, music, business, sports, politics). It features 61% person entities and 91.7% single-word mentions, linked to multiple knowledge graphs (DBpedia, YAGO, Wikidata), making it ideal for precision-focused homonym resolution testing.

Table 2: Statistics of the entity linking benchmarks used for evaluation.

Benchmark	Samples	Labels	Named Entities	Unknown	Ontology
AIDA-CoNLL	231	5616	4473	1132	Wiki
KORE50	50	144	140	1	Wiki
MSNBC	20	755	656	89	Wiki
News-Fair v2.0	120	1435	1018	169	Wiki
BC2GN	262	3223	793	282	NCBIGene
NCBI	100	960	202	92	MESH
MedHistory	52	235	145	1	HumanDO
MedHistory-hs	32	54	40	0	HumanDO

MSNBC contains 20 news articles from the MSNBC website with 755 mentions linked to Wikipedia, including 89 unknown entities (12%). Its contemporary news content and moderate multi-word mention distribution (43%) make it suitable for evaluating systems on incomplete knowledge base coverage and emerging entities in rapidly evolving domains.

News-Fair v2.0 provides 120 randomly-sampled news articles with annotated mentions linked to Wikidata, addressing biases in older benchmarks. Created through systematic annotation rules, it includes non-named entities and diverse topics, offering a realistic, balanced evaluation environment with reduced knowledge base coverage issues.

BC2GN (BioCreative II Gene Normalization) is a gene entity linking corpus from PubMed abstracts, originally limited to human genes but re-annotated at mention-level through GNormPlus for multi-species coverage. Linked to the massive NCBI Gene ontology (42M+ entities, 47.37% homonyms), it features short gene mentions (62% with numerals) requiring species-specific disambiguation.

NCBI (NCBI Disease Corpus) contains 960 disease mentions normalized to CTD Diseases (MEDIC). It features 15.62% unseen entities and 19.27% unseen synonyms, with abundant abbreviations requiring contextual disambiguation, serving as the standard public domain disease normalization benchmark.

D Full Experiments

Table 3: Full results of Evaluation of OneNet and CbEL. MedHistory-hs is a sub-dataset of MedHistory with terms flagged by experts as historical.)

Benchmark	CbEL			Onenet		
	NER F1	Disamb F1	Full F1	NER F1	Disamb F1	Full F1
Common EL Benchmarks						
aida-conll	0.646	0.510	0.330	0.664	0.292	0.194
kore50	0.829	0.526	0.437	0.834	0.181	0.151
msnbc	0.659	0.688	0.453	0.652	0.406	0.264
news-fair-v2	0.618	0.663	0.395	0.646	0.440	0.270
Medical Benchmarks						
BC2GN	0.394	0.174	0.064	0.381	0.110	0.034
NCBI	0.668	0.778	0.422	0.537	0.780	0.437
MedHistory	0.294	0.500	0.140	0.192	0.492	0.123
MedHistory-hs	0.194	0.785	0.152	0.114	0.736	0.084

Confidence-based re-linking with enhanced context drives performance gains. On NCBI, 90% of low-confidence entities gain higher confidence after the first loop, dropping to 46% in the second loop. This indicates that most entities are successfully re-considered and re-linked in the first iteration, yielding

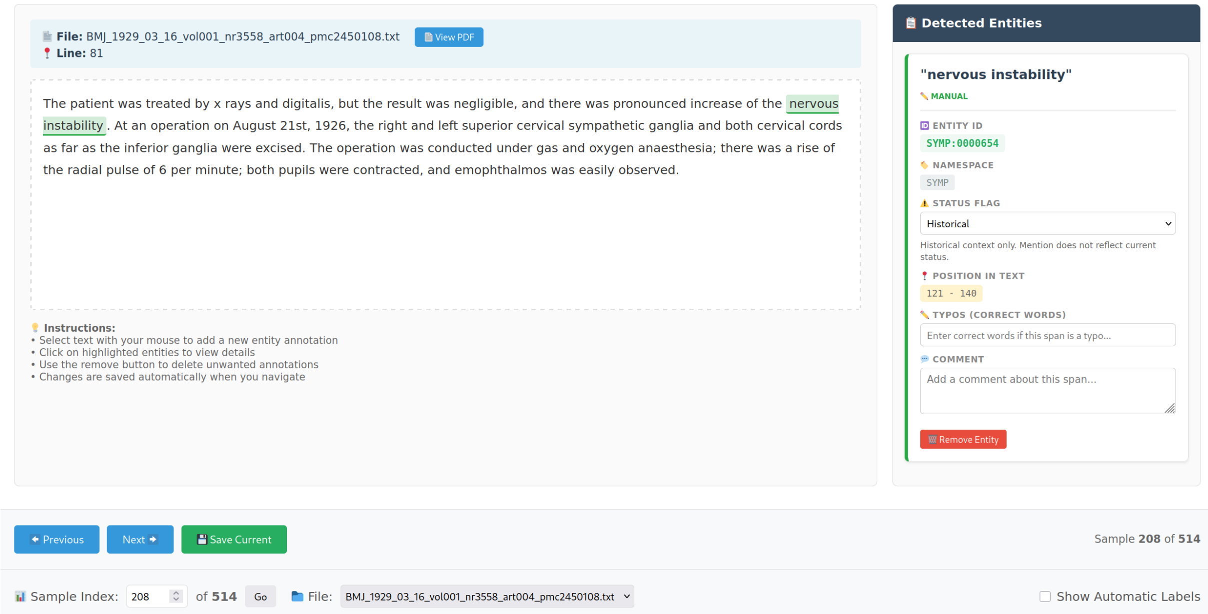

Figure 3: The annotation interface of our web-based annotation tool.

consistent results that require no further refinement. This validates the value of sequential confidence-based improvements for LLM-based entity linking.