Artificial Intelligence Review (2025) 58:140
https://doi.org/10.1007/s10462-024-11102-9

Musical heritage historical entity linking

Arianna Graciotti[1] · Nicolas Lazzari[1,2] · Valentina Presutti[1] · Rocco Tripodi[3]

Accepted: 29 December 2024 / Published online: 20 February 2025
© The Author(s) 2025

Abstract
Linking named entities occurring in text to their corresponding entity in a Knowledge Base (KB) is challenging, especially when dealing with historical texts. In this work, we introduce Musical Heritage named Entities Recognition, Classification and Linking (MHERCL), a novel benchmark consisting of manually annotated sentences extrapolated from historical periodicals of the music domain. MHERCL contains named entities under-represented or absent in the most famous KBs. We experiment with several State-of-the-Art models on the Entity Linking (EL) task and show that MHERCL is a challenging dataset for all of them. We propose a novel unsupervised EL model and a method to extend supervised entity linkers by using Knowledge Graphs (KGs) to tackle the main difficulties posed by historical documents. Our experiments reveal that relying on unsupervised techniques and improving models with logical constraints based on KGs and heuristics to predict `NIL` entities (entities not represented in the KB of reference) results in better EL performance on historical documents.

Keywords Historical documents · Named entity recognition · Named entity classification · Entity linking

The authors are listed in alphabetical order.

✉ Arianna Graciotti
arianna.graciotti@unibo.it

Nicolas Lazzari
nicolas.lazzari3@unibo.it

Valentina Presutti
valentina.presutti@unibo.it

Rocco Tripodi
rocco.tripodi@unive.it

[1] LILEC, University of Bologna, Bologna, Italy

[2] Computer Science Department, University of Pisa, Pisa, Italy

[3] DAIS, Ca' Foscari University of Venice, Venice, Italy

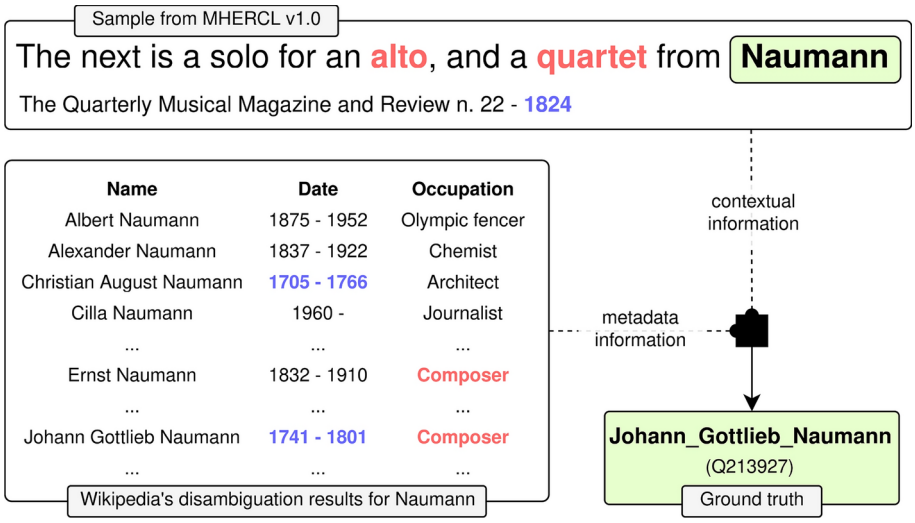

Fig. 1 Example sentence from an 1824 document in the Polifonia Corpus, with *Naumann* as the entity to be linked. The figure includes a selection of results from the Wikipedia disambiguation page (https://en.wikipedia.org/wiki/Naumann). Using sentence context and metadata from both the entity and the document allows for identifying the plausible entity, which in this case is unique

1 Introduction

Named Entity Recognition (NER), Named Entity Classification (NEC), and Entity Linking (EL)—also referred to as Entity Disambiguation (Bunescu et al. 2006)—are essential tasks in Knowledge Extraction (KE). NER involves identifying spans of text (*mentions*) that refer to named entities, such as people, locations, or organizations. NEC then assigns a specific entity type to each mention according to a predefined set of categories. EL links each mention to a unique entity within a reference Knowledge Base (KB), such as Wikipedia, DBPedia, or Wikidata. This linking process considers the context in which the mention occurs and, in some cases, information from the NER and NEC tasks (Tedeschi et al. 2021).

The extensive digitization of historical documents[1] is fundamental to the study and preservation of cultural heritage. The automatic processing of these documents revealed new challenges for NER, NEC, and EL models, deriving from the noisy quality of digitized historical text resulting from the use of Optical Character Recognition (OCR) technology or from variations between contemporary and historical language, such as differences in spelling, sentence structure, or changes in entities names over time.

In addition to these challenges, the scarcity of annotated resources poses a significant barrier to Historical Entity Linking (HEL). State-of-the-art (SotA) EL models, trained on contemporary web-based datasets, perform poorly on historical texts, where entities occur infrequently or are absent in the training data (Hoffart et al. 2011; Chen et al. 2021; Möller et al. 2022). For instance, consider the sentence "The next is a solo for an alto, and a quartet from Naumann" in Fig. 1, extracted from the 1824 issue of the historical music periodical

[1] We define historical documents as any textual materials produced or published up to 1979, as done by Ehrmann et al. (2023).

The Quarterly Musical Magazine And Review, part of the Polifonia Corpus[2]. The mention of *Naumann* is challenging to resolve: the Wikipedia disambiguation page lists over 30 possible entities,[3] each appearing infrequently (0 to 3 times) in typical training sets (Wu et al. 2020). Understanding the context of the mention aids in linking it to the correct entity. Music-specific terms in the sentence suggest a person relevant to the music domain. Knowing the date of the source document allows additional filtering, as entities born after 1824 shall be excluded. SotA models that ignore such contextual and temporal cues struggle with HEL, which is inherently more complex than linking entities in contemporary texts.

Furthermore, entity linkers tend to exhibit *popularity bias* (Kandpal et al. 2023), favouring frequently occurring entities in their training sets over less common (*long-tail*) ones. This bias is particularly problematic in historical documents, where entities are often less prominent than contemporary ones and appear less frequently in widely used KBs such as Wikipedia.

Finally, historical documents often contain named entity mentions without corresponding entities in a KB, known as `NIL` links (Sevgili et al. 2022; Guellil et al. 2024). Even though explicitly handling `NIL` links is fundamental in HEL, most SotA entity linkers lack this capability.

Following these arguments, our contribution is threefold:

1. We release Musical Heritage Historical named Entities Recognition, Classification and Linking (MHERCL), a gold-standard resource of historical documents focused on musical heritage and annotated for NER, NEC and EL. Although domain-specific, MHERCL aligns with typical HEL resources, which are mainly derived from newspapers (Ehrmann et al. 2022b) and is composed of a high number of long-tail entities with respect to their popularity.
2. We propose Entity Linking Dynamics (ELD), an unsupervised EL model based on a game-theoretical approach where mentions and entities interact to reach a consensus on the correct link. This approach addresses popularity bias by avoiding a supervised training phase, enabling the model to perform well on low-frequency entities often present in historical texts. Using an unsupervised approach, we bypass the limitations of labelled data, which may not cover all possible historical variations and interpretations of an entity, leading to potentially stronger generalization in scenarios where labelled data is sparse or incomplete.
3. We introduce a method that leverages constraints extracted from a Knowledge Graph (KG) to reduce implausible linking in retrieval-based EL models. We extend (Wu et al. 2020, BLINK) and propose Constrained-BLINK (C-BLINK), a version that incorporates these constraints, allowing for improved performance on historical texts. Moreover, we systematically explore how C-BLINK can be extended to account for `NIL` links, which are widely found in historical documents.The code for our experiments is publicly available on GitHub[4], and the dataset is also available on HuggingFace[5].

[2] The Polifonia Corpus is a multilingual and diachronic corpus focused on Musical Heritage. Available at https://github.com/polifonia-project/Polifonia-Corpus.

[3] Available at https://en.wikipedia.org/wiki/Naumann, last accessed May 22, 2024.

[4] Available at the URL https://github.com/polifonia-project/historical-entity-linking.

[5] Available at https://huggingface.co/datasets/n28div/MHERCL.

This paper is organised as follows: Sect. 2 reviews key contributions on EL benchmarks and models, focusing on those addressing the specific challenges of historical documents. Section 3 describes the documents selected for creating MHERCL and provides details on annotation guidelines and dataset statistics. In Sect. 4, we outline the implementation of ELD and describe how we adapted BLINK with typological and temporal constraints to create C-BLINK. Section 5 presents experimental results. Finally, Sect. 6 summarizes our contributions and suggests directions for future research.

2 Related works

In this section, we examine relevant datasets and benchmarks (Sect. 2.1) as well as models (Sect. 2.2) pertaining to EL. We present a detailed analysis of both general-purpose and historical document-specific benchmarks and methodologies.

2.1 Benchmarks

2.1.1 General purpose

Recent surveys indicate that historical texts are rarely utilized in datasets for training, fine-tuning, or evaluating EL models (Möller et al. 2022; Guellil et al. 2024). The most commonly used EL datasets for the English language include AIDA-CONLL (Hoffart et al. 2011), which contains Reuters news stories from 1996 to 1997; AQUAINT (Graff 2002), a derivative of AIDA-CONLL; MSNBC (English), which is made of online news from 2007; ACE2004 (Mitchell et al. 2005), which is composed of web-collected news articles (also in Chinese and Arabic, other than in English); CWEB (Gabrilovich et al. 2013), which consists of Wikipedia and generic web pages; WIKI (Alani et al. 2018), which is made of Wikipedia pages; TAC (Ji et al. 2010), which contains Wikipedia pages extracted in 2008; and KORE50 (Hoffart et al. 2012), which is composed of web documents with a focus on highly ambiguous entities.

Several studies have targeted particularly challenging EL scenarios. The ShadowLink dataset (Provatorova et al. 2021) includes 16,000 short text snippets from websites in English, annotated with highly ambiguous entities. This dataset focuses on named entities with identical surface forms linked to common and rare entities, causing the former to overshadow the latter. The TempEL dataset (Zaporojets et al. 2022) offers time-stratified English Wikipedia snapshots from 2013 to 2022, enabling the evaluation of EL models on both continuous and newly emerging entities. Additionally, Benkhedda et al. (2024) concentrated on the English language Community-generated digital content (CGDC) for digital archives and developed a dataset of textual metadata linked to Wikidata information. Although this metadata includes historical named entities due to its intended use (enhancing documentation for historical digital archives), it remains a product of contemporary digital text and does not exhibit the distinctive features of historical documents.

Zhu et al. (2023) focuses on entity mentions that cannot be resolved against the reference KB and introduces NEL, an annotated dataset of English Wikipedia excerpts with a significant percentage of `NIL`, i.e. entities that do not exist in the KB of reference. The significant presence of `NIL` entities is particularly relevant also for HEL, as historical docu-

ments present a high occurrence of out-of-KB entities. Despite addressing `NIL` links, this work concentrated on digitally-born texts, which display different linguistic characteristics compared to digitalized documents.

2.1.2 Historical

Ehrmann et al. (2023) reviews resources for historical NER. Of particular relevance for the work described in this paper are the resources collected for the HIPE-2020 (Ehrmann et al. 2020b, a) and HIPE-2022 (Ehrmann et al. 2022a) evaluation campaigns, as they include EL annotations. In particular, the HIPE-2022 corpus comprises six datasets spanning English, Finnish, French, German, and Swedish. The datasets are composed by collecting historical newspapers and classic commentaries of 200 years time-span sourced from: (i) (Hamdi et al. 2021), which includes historical newspaper articles in French, German, Finnish, and Swedish dating back to 19C–20C; (ii) *SoNAR* (Menzel et al. 2021), which includes historical newspaper articles from the Berlin State Library newspaper collections in German (19C–20C); (iii) *Le Temps* (Ehrmann et al. 2016), which contains historical newspaper articles from two Swiss newspapers in French (19C–20C); (iv) *Living with Machines* (Ehrmann et al. 2016), which contains *TopRes19th*, a dataset of historical newspaper articles from the British Library newspapers in English (18C–19C), whose annotations focus on toponyms documents; (v) *AjMC*, which contains annotated *Classical Commentaries* (19C). We employ the HIPE-2020 English test set in our experiments, as detailed in Sect. 5.

Another effort dedicated to historical documents brought to the creation of Giorgio Vasari's *Lives of The Artists* (Santini et al. 2022), which contains NER, NEC and EL annotations towards Wikidata. A similar initiative is followed by Blouin et al. (2024), which carries out NER and EL towards Wikidata annotations on a historical newspaper in the Chinese language published between 1872 and 1949 and on bilingual (Chinese-English) biographies dating back to the first half of the 20C.

Our resource enriches the field of HEL, which suffers from a scarcity of resources. It addresses the music domain, a sector not yet covered by existing HEL resources. Moreover, while contemporary EL benchmarks often emphasize English, HEL resources are not predominantly in English. For example, our dataset contains almost five times as many annotated named entities as the HIPE-2020 English test set, positioning itself as a valuable complement to existing resources for NER, NEC and EL on historical documents in English.

2.2 Models

2.2.1 General purpose

Current research on EL models predominantly employs Pre-trained Language Models (PLM). Two primary approaches have been identified: retrieval-based and generative methods. Retrieval-based methods involve encoding mentions into a dense representation, which is then compared with a similarly dense representation calculated from the documents in a KB, such as Wikipedia. An optional re-ranking phase can merge these dense representations for improved accuracy. This approach is exemplified by tools like van Hulst et al. (2020, REL), which features a modular architecture of neural components, as well as BLINK and its multilingual extension (Plekhanov et al. 2023, BELA).

A similar approach is proposed in Zhang et al. (2022, EntQA) and Orlando et al. (2024, ReLiK). Unlike in BLINK, the linking phase is framed as a Question Answering (QA) problem. A plausible set of entities is first retrieved using a dense representation similar to the one from BLINK. The difference is that the named entity is not recognized beforehand, but the model jointly recognizes mentions and possible entities that might be linked. The correct entity is then chosen based on the context of the sentence.

Christmann et al. (2022, CLOCQ) is a system designed to extend beyond the task of linking named entities to their corresponding entries in a knowledge base, addressing the broader challenge of complex question answering over knowledge bases. One of its core functionalities includes disambiguating named entities, making it relevant to our work.

A different approach, based on generative language models instead of classification architectures, is proposed in De Cao et al. (2021, GENRE). This approach retrieves entities by generating their unique names, left to right, conditioned on the context. Such an approach directly captures relations between context and entity name and reduces the memory footprint of performing queries on large search spaces. An extension of this model, with support for multiple languages, is proposed in De Cao et al. (2022, mGENRE). Barba et al. (2022, ExtEnD) builds upon the approach used by mGENRE, but reformulates it by framing that as a text extraction task. The set of generated candidates is concatenated to the span, thereby overcoming GENRE's limitation of being unable to see the candidate entities from which to choose.

EL models that leverage symbolic knowledge from KBs can significantly enhance the plausibility of linking entities. As demonstrated by Tedeschi et al. (2021), it is possible to outperform traditional methods and minimize confusion during the linking process by filtering the pool of linkable entities based on the properties of the mention. This method involves classifying named entity mentions into fine-grained categories, ensuring that potentially linkable entities share the same type as the mention. Consequently, the model focuses on determining the most likely entity rather than learning the plausibility of each entity implicitly.

Building on similar assumptions, Ayoola et al. (2022, ReFinED) focuses on optimizing performances on long-tail entities by leveraging symbolic knowledge extrapolated from KBs (entities descriptions', types, and inter-entities relations in context). Given its focus on long-tail entities, ReFinED encompasses the management of out-of-KB entities by including the `NIL` label as a possible candidate for the linking step.

Except for ReFiNeD, SoTA entity linkers presuppose a correct link for each mention is always present in the KB of reference. As a result, impaired focus has been dedicated by the research community to the `NIL` prediction issue. Different approaches have been proposed to predict whether a `NIL` prediction should be inferred from the similarity score. This includes heuristic methods, such as a threshold on the score, or machine-learning-based methods, such as training a binary classifier (Sevgili et al. 2022). To address this issue, Dong et al. (2023) propose BLINKout, a method that uses a dynamic `NIL` representation and classification approach, incorporating synonyms and built on BERT-based EL.

2.2.2 Historical

The scarcity of established resources and benchmarks for HEL results in fewer solutions than conventional EL. Agarwal et al. (2018) pioneered the research in this field by propos-

ing diaNED, a time-aware entity disambiguation approach for diachronic corpora, which leverages KB-driven temporal information and text-driven temporal expressions.

The HIPE-2022 evaluation campaign collates the most recent efforts in implementing models for NER, NEC and EL on historical documents. The highest scoring model for the challenge on CLEF-HIPE-2022 English language data for EL was proposed by the L3i team (Boros et al. 2022), which builds upon the same multilingual model based on a BiLSTM architecture (Kolitsas et al. 2018) that the team presented when they introduced it for CLEF-HIPE-2020 (Boros et al. 2020). Their model is enhanced with a filtering process (Pontes et al. 2022) to disambiguate historical references using the typological and temporal information in Wikidata. All the models that took part in the evaluation campaign implemented strategies to take into consideration the prediction of the `NIL` link.

In this paper, we introduce C-BLINK, which enriches the scenario of the aforementioned related works by extending BLINK. Our model leverages time- and type-related knowledge from our chosen KB of reference, Wikidata, to refine the candidate generation process. After filtering out the candidates that fail the plausibility criteria scoped by the time- and type-related constraints, we test various methods, including threshold-based and machine learning-based approaches, to identify `NIL` links. The original work does not cover both aspects.

Also, we propose ELD, which differs from existing approaches in HEL by framing the EL problem using a game-theoretical approach. Named entity mentions, their surrounding context, and KB entities are embedded into dense vectors. The resulting vectors are used to compute pairwise similarities. By playing against each other, named entities have to collaborate with most similar entities, eventually converging to the final link decision. The model is unsupervised.

3 MHERCL: a new benchmark for HEL

This section describes MHERCL, the new benchmark we contribute for NER, NEC and EL on historical documents. We describe how we build the sample of sentences, the annotation guidelines, and the Inter-Annotator Agreement (IAA) alongside the benchmark statistics.

3.1 Sampling

MHERCL contains manually annotated sentences extrapolated from the Polifonia Corpus[1]. The Polifonia Corpus is a diachronic and multilingual corpus specialized in the music domain. It is composed of multiple modules, which contain documents from various data sources. A subset of the *Periodicals* module, containing documents whose publication dates range from 1823 to 1900, is used as the set of sentences within MHERCL. The subset is selected such that sentences are in English and contain at least one named entity. Because the Polifonia Corpus's focus is on music, it provides a unique domain-specific perspective on the EL problem. Nonetheless, the documents within MHERCL align closely with the prevailing document types already utilized in existing benchmarks within the field, namely historical newspapers (Ehrmann et al. 2022b).

3.2 Data annotation

The sentences included within the selected subset are manually annotated by two annotators selected from Foreign Languages and Literature undergraduate students at the University of Bologna. Both annotators have been trained on the NER, NEC and EL annotation tasks.

The annotation work was performed under the criteria that a named entity is a real-world thing indicating a unique individual through a proper noun (Jurafsky and Martin 2023). The annotation process involves inspecting the sentences and identifying the named entities, eventually linking them to their corresponding Wikidata unique identifier (QID).

Named entities are recognized, classified, and linked following the Abstract Meaning Representation [AMR (Banarescu et al. 2013)] named entity annotations guidelines.[6] The guidelines report instructions on how to classify the named entities according to a predefined list of types.

A full description of the types used for NEC in MHERCL is provided in Appendix C. Annotation guidelines are extensively described in Appendix A.

3.3 Inter-annotator agreement

Inter-annotator agreement (IAA) measures the reliability of human annotations by assessing the consistency among annotators. To evaluate IAA, two annotators independently annotated the same 101 sentences extrapolated from MHERCL. We computed Krippendorff's alpha (Hayes and Krippendorff 2007; Castro 2017) for nominal metrics on the resulting annotations. We chose Krippendorff's alpha due to its flexibility and resilience in handling missing values, as demonstrated in other contexts (Tripodi et al. 2022). Table 1 presents the number of sentences and tokens of the IAA sample.

With an annotator agreement higher than 0.8, MHERCL's annotation quality is reliable (Hayes and Krippendorff 2007) and can be treated as a gold standard.

When we calculate the IAA, we identify instances where the annotations of the two annotators do not match, as shown in Table 1. We observe that this discrepancy is often due to the complexity of the HEL tasks. For example, historical names used when the periodicals were issued might not be used anymore in current times.

Consider the following example sentence from MHERCL, extrapolated from an issue of the music periodical *The Harmonicon*, dating back to 1833:

Example 1 I find the following sensible critical remarks among my papers; they were copied from a weekly work (the Court Magazine, I think)

Table 1 IAA within MHERCL

#tokens (Annotated)				
#sents	#tokens (Tot.)	#matching	#unmatching	Krippendorf's alpha
101	6589	656	124	0.82

[6] Available at the URL https://github.com/amrisi/amr-guidelines/blob/master/amr.md#named-entities, last time accessed on August 12th, 2024.

In Example 1, one annotator linked the mention "Court Magazine" to "La Belle Assemblée"[7] while the other annotator labeled it as NIL. As reported in Wikipedia, "Court Magazine" resulted from the historical merging of different periodicals, including "La Belle Assemblée":

> The magazine was published as La Belle Assemblée from 1806 until May 1832. It became The Court Magazine and Belle Assemblée from 1832 to 1837. After 1837, the Belle Assemblée name was dropped when the magazine merged with the Lady's Magazine and Museum (itself a merger of The Lady's Magazine and a competitor) to become The Court Magazine and Monthly Critic.

Nevertheless, the correct annotation is NIL since "Court Magazine" and "La Belle Assemblée" refer to distinct entities. This case illustrates how such ambiguities can confuse annotators.Consider another example sentence from MHERCL, extrapolated from an issue of the music periodical *The Quarterly Musical Magazine And Review*, dating back to 1828:

Example 2 Sommer, organist to the Court of Vienna; this is the greatest compliment which can be paid to his talents.

In Example 2, one annotator labeled "Court of Vienna" as NIL, while the other linked it to the "Vienna State Opera".[8] The source of this ambiguity can be traced back again to information present in Wikipedia, which misled the annotator:

> The Vienna State Opera [...] is an opera house and opera company based in Vienna, Austria [...]. The opera house was inaugurated as the "Vienna Court Opera" (Wiener Hofoper) in the presence of Emperor Franz Joseph I and Empress Elisabeth of Austria. It became known by its current name after establishing the First Austrian Republic in 1921.

After its construction, the opera house was known as the "Vienna Court Opera". However, the construction happened after the date on which the periodical from which Example 2 is extracted was published, which is 1828, as it can be read in MHERCL's metadata. Therefore, the correct annotation is NIL.After calculating the IAA and identifying the discrepancy cases, we reconciled the output by manually aligning each problematic instance with the Annotation Guidelines. This ensures that the final version of the benchmark reports the correct cases.

3.4 Statistics

Tables 2 and 3 report the main statistics for MHERCL, and compare it to the most similar resource available for HEL, HIPE-2020. MHERCL is composed of 875 sentences, extrapolated from 76 historical periodicals. They include 1805 unique named entities belonging to 58 different types. As it is possible to observe from the comparison, MHERCL is larger

[7] Available at the URLs https://en.wikipedia.org/wiki/La_Belle_Assemblée and https://www.Wikidata.org/wiki/Q3206551, last time accessed on August 12th, 2024.

[8] Available at the URLs https://en.wikipedia.org/wiki/Vienna_State_Opera and https://www.Wikidata.org/wiki/Q209937, last time accessed on August 12th, 2024.

Table 2 Documents' statistics for MHERCL and HIPE-2020 test set

Dataset	#docs	#sents	#tokens	#tokens per sentence (average)
MHERCL	76	875	27,549	31.5
HIPE-2020	46	553	16,635	30

Table 3 Named entities statistics for MHERCL and HIPE-2020 test set

Dataset	Count	#mention	#types	Noisy	NIL
MHERCL	All	2370	*N/A*	0.14	0.30
	Unique	1805	58	*N/A*	0.38
HIPE-2020	All	449	*N/A*	0.05	0.40
	Unique	232	5	*N/A*	0.54

Table 4 Top 10 named entity types occurring in MHERCL

Type	#
PERSON	1253
CITY	262
MUSIC	187
ORGANIZATION	93
WORK-OF-ART	85
COUNTRY	80
BUILDING	52
OPERA	52
THEATRE	42
WORSHIP-PLACE	41

than HIPE-2020 English test set with respect to the number of sentences included and the annotated entities. Also, it is more varied regarding source documents and the named entity types considered.

As reported in Table 3, 30% of the total of annotated named entity mentions could not be linked to any Wikidata entity, which resulted in NIL annotations. This particular aspect makes MHERCL a challenging dataset from a linking standpoint, as the model must be able to predict an entity conservatively—i.e., predicting NIL—or it is impossible to obtain an accuracy higher than 70%. Moreover, in Table 3, the amount of noisy entities is reported. These are the annotations impacted by errors due to OCR procedure on the original periodicals.

Table 4 reports the top 10 named entity types occurring in MHERCL in order of frequency. The types distribution reveals the benchmark's specificity to the music domain due to the presence of domain-specific types such as MUSIC, OPERA, and WORK-OF-ART. A full description of the types used for NEs classification is provided in Appendix C.

3.5 Popularity study

Numerous studies tackle popularity bias in knowledge-intensive downstream tasks (Kandpal et al. 2023; Mallen et al. 2023; Sun et al. 2024), including EL (Ilievski et al. 2018; Ayoola et al. 2022). To show the extent to which MHERCL may contribute to the advancements of research for EL on long-tail (i.e., low popularity) entities, we analyze the popularity distribution of its named entities, and we compare it to: (1) the HIPE-2020 English test set (Ehrmann et al. 2020a, b); (2) the AIDA CoNLL-YAGO EL benchmark (English)

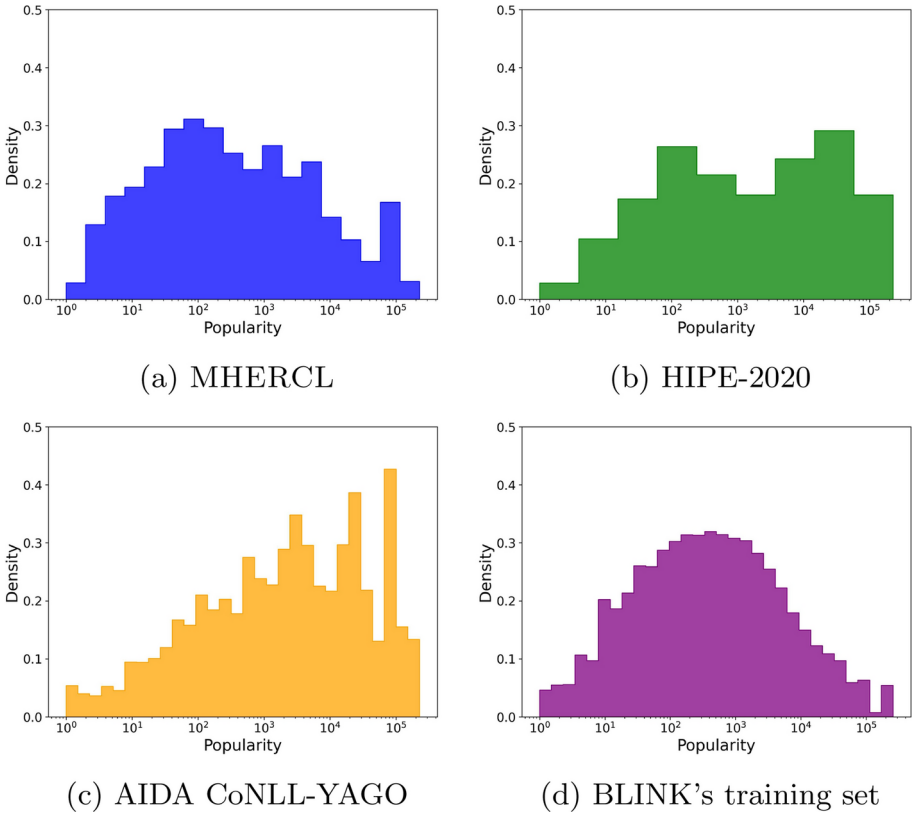

Fig. 2 Figures showing the distribution of named entity popularity in MHERCL (**a**), HIPE-2020 (**b**), AIDA-COnLL-YAGO (**c**) benchmarks and BLINK's training dataset (**d**). Each figure contains a histogram showing the density of entities' popularity levels: the higher the density, the more common the entities with the corresponding popularity value. Popularity is computed as the frequency of occurrence of each named entity's QID as an internal link in Wikipedia. NIL are excluded in MHERCL and HIPE-2020

(Hoffart et al. 2011), (3) BLINK's Training Set (TS), which consists of a 9 million examples subset of the May 2019 English Wikipedia dump. AIDA CoNLL-YAGO and BLINK's TS are taken in their KILT (Petroni et al. 2021) release. As a proxy for popularity, we calculate each named entity's QID frequency of occurrence as an internal link in Wikipedia[9]. NIL entities in MHERCL and HIPE-2020 are excluded from the analysis.

Aligning with Mallen et al. (2023), we illustrate in Fig. 2 the distribution of named entities' popularity across the three datasets chosen. As can be observed in Fig. 2a, MHERCL displays higher density levels at lower popularity values. This means that, on average, an entity randomly drawn from MHERCL is more likely to be associated with low popularity values (between 10 and 10^3). On the contrary, as shown by Fig. 2c, AIDA CoNLL-YAGO named entities exhibit a higher density at higher popularity values, showing that traditional benchmarks mostly focus on highly represented entities. Despite its historical focus, as shown in Fig. 2b, HIPE-2020 presents a distribution that resembles the sum of two normal distributions centred at moderately popular entities ($\approx 10^2$) and highly popular entities ($\geq 10^4$). It

[9] For popularity calculation, we used the Wikipedia dump enwiki-20220120 dump.

is hence biased towards popular entities, similarly to AIDA CoNLL-YAGO, but also maintains a considerable representation of moderately popular entities, which are the prevalent type of entities in training datasets (see BLINK's TS distribution of Fig. 2d). Indeed, from Fig. 2d, it can be seen how BLINK's TS, which is derived from Wikipedia, has a wide coverage of highly popular entities as well as low-popular ones. This proves that models processing historical documents can benefit from training on general-purpose datasets, as their popularity distribution includes low-popularity entities as well.

As can be observed in Fig. 3 reporting the Kernel Density Estimation (KDE) plots, all three datasets share a common distribution of entities within the mid-range popularity values (10^1 to 10^4). Despite this overlap, the plot underlines that HIPE-2020 and MHERCL are complementary datasets for historical documents since they focus on a similar domain but display different data distributions. From a popularity point of view, HIPE-2020 is more similar to AIDA CoNLL-YAGO, while MHERCL characterizes itself as a benchmark complementary to standard ones—e.g. AIDA CoNLL-YAGO. This further highlights the novelty and importance of our dataset for studies focusing on low-popularity entities, extending its value beyond HEL to broader applications in EL research.

4 Models

Due to the limitations of current entity linkers in HEL discussed in Sect. 1, models tailored to the unique requirements of historical texts are needed. In this section, we present two methods based on two different architectures to enhance entity linkers for HEL.

The first method, Entity Linking Dynamics (ELD, Sect. 4.1), is an unsupervised model based on game-theoretic principles. It bypasses explicit training on labelled sentences, making it suitable for low-popularity entities typical of historical documents, and explicitly handles `NIL` predictions. The second method, Constrained-BLINK (C-BLINK, Sect. 4.2) extends BLINK by adding type and temporal plausibility constraints during the candidate retrieval phase and by including `NIL` prediction without the need to re-train the model.

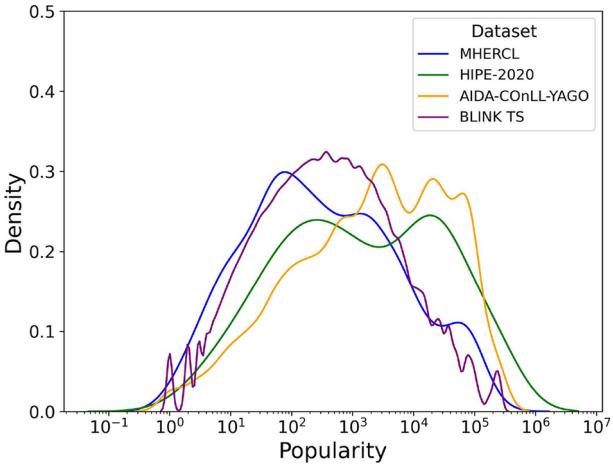

Fig. 3 Kernel Density Estimation (KDE) plot comparing the smoothed density functions of named entities popularity across the MHERCL, HIPE-2020, AIDA CoNLL-YAGO and BLINK's Training Set (TS) datasets. For MHERCL and HIPE-2020, only entities with a valid QID are considered, `NIL` entities being excluded from the analysis

4.1 Entity linking dynamics (ELD)

4.1.1 Theoretical foundation

ELD's approach to EL assumes that the disambiguation process underlying this task has to be coherent with the context in which entities appear, respecting time and logical constraints. It is grounded on relaxation labelling processes (RLP) (Hummel and Zucker 1983). This class of algorithms produce consistent labelling of the data thanks to the use of constraints among labels (entities in the context of EL) designed to make the label assigned to an object consistent with the labels assigned to the other objects in the same batch (sentence or paragraph in the context of EL). RLPs have been used successfully in different fields, including computer vision (Pelillo 1997), Natural Language Processing (NLP) (Tripodi and Navigli 2019), and graph processing (Rota Bulò et al. 2011). One possible formulation of RLPs is in terms of game theory (Miller and Zucker 1991). In this formulation, objects to be labelled are interpreted as the players of a non-cooperative game (Nash 1951), and the constraints on labels are the strategies that the players can select to play the game.

In game theory, particularly in non-cooperative games, the players are considered rational and select the strategy that maximizes a predefined payoff. The selection is performed considering the strategies a co-player can select in response to a strategy. The payoff of different strategy combinations is stored in a payoff matrix that defines the game. For example, consider a two-player game, i.e., *matching pennies*, where a coin is flipped, and the strategies of the players correspond to the selection of one face of the coin (head or tail). Player 1 wins if both players select the same face of the coin, while player 2 wins if they choose different faces. If both players select heads, player 1 will have a positive payoff, while player 2 will have a negative payoff. The payoff matrix, reported in Table 5, defines the payoff of both players with respect to their gaming strategy—i.e. if they chose heads or tails. As a consequence, the outcome of a game does not depend on the strategy selected by a single player but on the combination of the strategies selected by the players involved in the game. An equilibrium state for the game described above, i.e. the optimal strategy that both players should follow, is to use a mixed strategy in which heads and tails are played with the same probability. In this way, it is difficult for a player to guess the strategy that the other player will use. In this case, we say that the players are using *mixed strategies*, which are probability distributions over pure strategies.

A mixed strategy for a player i is defined as a probability vector $\mathbf{x_i} = (x_1, \ldots, x_{m_i})$, such that $x_j \geq 0$ and $\sum x_j = 1$. Each mixed strategy corresponds to a point in the simplex Δ_m, whose corners correspond to pure strategies. The intuition is that players randomize over strategies according to the probabilities in $\mathbf{x_i}$. Each mixed strategy profile lives in the mixed strategy space of the game, given by the Cartesian product $\Theta = \Delta_{m_1} \times \Delta_{m_2} \times \cdots \times \Delta_{m_n}$.

Table 5 Payoff matrix for the game of *matching pennies*

		PLAYER 2			
		Heads		Tails	
PLAYER 1	Heads	1	−1	−1	1
	Tails	−1	1	1	−1

The number on the left of each cell of the matrix indicates the payoff obtained by player 1, and the number on the right indicates the payoff of player 2. Player 1 plays according to the rows of the matrix, and player 2 with the columns

In a *two-player game*, a strategy profile can be defined as a pair $(\mathrm{x}_i, \mathrm{x}_j)$. The expected payoff for this strategy profile is computed as:

$$u(\mathrm{x}_i, \mathrm{x}_j) = \mathrm{x}_i^T \cdot A_{ij}\mathrm{x}_j$$

where A_{ij} is the $m_i \times m_j$ payoff matrix between player *i* and *j*. In evolutionary game theory (Weibull 1997), the games are played repeatedly, and the players update their mixed strategy distributions over time until no player can improve the payoff obtained with the current mixed strategy. This situation corresponds to the equilibrium of the system. This dynamic process allows to perform consistent labeling of the data (Miller and Zucker 1991), and from this aspect, we assign a name to our system, i.e. Entity Linking Dynamics.

In ELD, the EL task is framed as a game in which each player is a mention that has to select an entity to be linked to. The strategy of the game corresponds to the entities to which a mention can be linked towards a KB. The Entity Linking games are formulated by modeling the ability of each player to select the most consistent strategy, i.e. the strategy that grants the highest payoff. This is done using a KB to select all the possible entities to which each mention can be associated. These entities are interpreted as the strategies that the player can select. The linking process is performed jointly with Word Sense Disambiguation (WSD), i.e., associating each content word in a sentence with its meaning as encoded in a KB such as WordNet (Miller 1995). This allows the exploitation of contextual information to perform consistent data labelling, disambiguate entities, and consider nouns and verbs around them. ELD leverages historical context as implicit knowledge, which helps the model to make connections between entities and assign a higher payoff to entities that are related. This allows for consistent labeling of the data, which means disambiguating an entity taking into account the relations that it has with all other entities in the text. Related entities could be entities that existed in the same period or entities that can be associated with the same domain, such as artists and musicians in the context of MERCHL. This approach enables the model to capture nuanced relationships that may not be explicitly defined in labelled training datasets, particularly when disambiguating entities individually rather than simultaneously, as in ELD.

4.1.2 ELD model

In ELD, an input sentence is pre-processed with a tokeniser, a part-of-speech tagger and a named entity recognition tagger to obtain a set of annotations $T = ((t_1, pos_1, ner_1), (t_2, pos_2, ner_2), \ldots, (t_n, pos_n, ner_n))$. Additionally, the text is fed into a Pre-trained Language Model to obtain a contextualised representation of each token $E = (\mathrm{e}_1, \mathrm{e}_2, \ldots, \mathrm{e}_n)$. The subtoken representations of *E* are averaged to match the tokenisation in *T*. The result is a feature vector for each token in *T*, which results in a $n \times w$ matrix *W* in which only the *n* nouns, verbs, adjectives, adverbs and named entities in *T* are maintained and *w* refers to the dimension of the feature vectors. These *n* tokens are the players of the game implemented in ELD. Based on the pairwise similarities computed using *W*, ELD weights each player's influence on each other such that similar players have a reciprocal influence on their strategy selection. For example, in the sentence:

Abe Lincoln played the guitar in the Dinosaurs band for 20 years.

the word *guitar* influences the mention *Abe Lincoln* in the entity *Abe_Lincoln_(guitarist)* rather than *Abe_Lincoln_(politician)*. This information is obtained by constructing a payoff matrix that encodes the similarity between the word *guitar* and the entities *Abe_Lincoln_(guitarist)* and *Abe_Lincoln_(politician)*.

The candidate generation phase is performed by selecting the possible senses/entities to which a word/mention can refer. We rely on two different KBs: WordNet for the selection of the senses of nouns, verbs, adjectives and adverbs, and Wikidata for the selection of entities connected to a mention marked as named entity.

In ELD, the strategy space is a probability distribution encoded in an $n \times m$ matrix X where n is the number of players (the mentions in the sentence) and m the entities to link to. It can be defined using a prior, if available, or through a uniform distribution. The payoff matrix is defined by computing embeddings of each entity and sense such that they all lie in the same latent space—i.e. they are compatible representations. In particular, sense embeddings are obtained using Ares embeddings (Scarlini et al. 2020), employing a multilingual language model to embed text annotated with WSD. In contrast, entity embeddings are obtained using the same multilingual language model used by Ares, i.e., mBERT (Devlin et al. 2019). The $m \times m$ matrix Z, represents the payoff matrix of the game. It is obtained by computing the pairwise similarity between entity embeddings. This formulation allows to guide the disambiguation dynamics towards the senses of an entity that are more similar to it. In the example above, the embedding *guitar* has higher contextual similarity to the entity *Abe_Lincoln_(guitarist)* than to other possible entities.

ELD exploits the matrix Z by relying on the replicator dynamic equation (Taylor and Jonker 1978) where the payoff of strategy h for player i is calculated as:

$$u(x_i^h) = x_i^h \cdot \sum_{j=1}^{n_i} (A_{ij} Z \mathrm{x}_j)^h \tag{1}$$

where x_i^h is the probability of strategy h for player i, according to the matrix X, x_j is the embedding of the entity j and n_i are the neighbours of player i, defined by the adjacency matrix A. Informally, A encodes which senses are neighbours of the mention i in the input sentence. The entity with a higher payoff with the target mention is selected for the final link.

4.2 C-BLINK

Constrained-BLINK (C-BLINK) extends BLINK by filtering entities based on type and temporal plausibility during candidate retrieval. Type and date information are sourced from an external KB (Wikidata).

In HEL, a *plausible* candidate is one whose type loosely aligns with the manually annotated named entity type and whose time information precedes the document's date. For instance, consider the following sentence from MHERCL, originally found in a 1824 issue of *The Quarterly Musical Magazine And Review*:

> Fabio Constantini flourished about the year 1630, and ultimately became maestro at the chapel of Loretto.

BLINK links *Fabio Constantini* to *Claudio Constantini*,[10] which is implausible, as the document dates to 1824 and *Claudio Constantini* was born in 1983.

Following the approach of Tedeschi et al. (2021), we limit the candidates considered by BLINK's to *plausible* entities. Specifically, C-BLINK extends BLINK's candidate retrieval by filtering its *bi-encoder* module using a collection of Boolean functions Φ. Each boolean function $\phi \in \Phi$ checks whether an entity *e* is plausible for a mention *m*. For example, it is possible to implement the approach of Tedeschi et al. (2021) by defining $\phi_t \in \Phi$ as that function that evaluates to true if the NEC class predicted for the mention *m* matches the class predicted for the entity *e*.

Moreover, we define $\phi_d \in \Phi$ as the function that evaluates to true only if the date associated with the entity *e* precedes the date associated with the mention *m*. In the example above, $\phi(\text{Fabio Constantini}, \text{Claudio Constantini})$ is false since *Fabio Constantini* is mentioned on a document issued in 1824 and *Claudio Constantini* was born in 1983.

In practice, the integration of such constraints in C-BLINK is implemented by directly annihilating the similarity scores computed by the bi-encoder for implausible entities. Our approach is agnostic to the choice of the plausibility function: $\phi \in \Phi$ can be heuristic, like ϕ_d or any arbitrary function, including a neural network. In the latter case, *m* and *s* are represented by their dense encodings from the bi-encoder. This allows for the straightforward reuse of pre-trained models without any additional training requirements.

The plausibility constraints avoid the prediction of logically inconsistent entities, but it does not directly address `NIL` predictions for entities that are not present in our reference KB. Since all the candidates considered by the model are plausible, we assume that if the correct entity is present in the KB, then it must be the one with the highest similarity with respect to the document's mention. Note that this assumption directly aligns with retrieval-based models since they are trained to maximise the similarity between the document's dense representation and its associated entity. Our `NIL` prediction approach leverages this assumption. If there is no document in the pool of candidates whose similarity is significantly higher than the rest with respect to the input mention (according to the heuristics described later), we infer the absence of a confident link and predict `NIL`.

As discussed in Sect. 2, different approaches have been proposed to perform `NIL` prediction. We design an extensive set of heuristics and classifiers to assess their effectiveness on this task, acting on three main aspects: (1) the candidates' scores, (2) the string similarity between the superficial mention found in the benchmarks and the predicted entity, (3) machine learning-based methods.

Table 6 describes the heuristics that act on the scores provided by the model. For the string similarity heuristics, we experiment with Levenshtein distance, Jaccard distance, and Hamming distance (Wang and Dong 2020). All those methods involve the use of a threshold value. For machine learning models, we experimented with Support Vector Machines (SVM), logistic regression, and decision trees.

We extensively test C-BLINK and report a fine-grained description of the effects of ϕ_d, ϕ_t and `NIL` heuristics on its performance in Sect. 5.

[10] Available at https://en.wikipedia.org/wiki/Claudio_Constantini and https://www.wikidata.org/wiki/Q5129347, last accessed on August 12th, 2024.

Table 6 Description of the statistical heuristics

Name	Condition	Description
Fixed threshold	$s_0 < \tau$	The top score should be higher than τ
Deviation from top	$\frac{s_0 - s_1}{(s_0+s_1)/2} < \tau$	The percentage difference between the two top scores should be higher than τ
Deviation from median	$\frac{s_0 - \overline{S}}{(s_0+\overline{S})/2} < \tau$	The percentage difference between the top score and the median of the scores ($\overline{S}$) should be higher than τ
Deviation from mean	$\frac{s_0 - \mu(S)}{(s_0+\mu(S))/2} < \tau$	The percentage difference between the top score and the mean of the scores ($\overline{S}$) should be higher than τ

$S = [s_0, \cdots, s_n]$ is used for the scores provided by the model to the entity that can be linked, $\tau \in [0, 1]$

5 Experiments

In this section, we detail our experimental setup designed to evaluate the performance of our HEL models. We conduct assessments of ELD and C-BLINK across both MHERCL and HIPE-2020, comparing them against established baselines. These baselines encompass publicly accessible retrieval-based and generative models: REL (van Hulst et al. 2020), BLINK (Wu et al. 2020), GENRE (De Cao et al. 2021), mGENRE (De Cao et al. 2022), ReFinED (Ayoola et al. 2022), ExtEnD (Barba et al. 2022), CLOCQ[11] (Christmann et al. 2022), and ReLiK (Orlando et al. 2024), as discussed in Sect. 2. Each model has been used following the instructions provided by the authors in their respective websites or repositories. Additionally, we experiment with Large Language Models (LLMs) and test the ability of GPT-4 o1-mini[12] and LLAMA 3.3 70B instruction tuned.[13] For both models, we rely on the same prompt instructing the model to produce the name of the Wikipedia page of a given named entity mention, or to answer with `NIL` if there is no relevant page. We report the prompt used in the appendix at Sect. E.

For C-BLINK and ELD, as anticipated in Sect. 4, we implement two plausibility functions addressing time and type constraints, detailed in Sects. 5.1 and 5.2 respectively. Additionally, we explore various approaches for `NIL` prediction within C-BLINK, employing both heuristic methods and machine learning classifiers, which are elaborated in Sect. 5.3.

5.1 Time constraint

The first criterion on which we base our candidate plausibility function is chronological consistency (time plausibility). We define $\phi_d \in \Phi$, the function that implements the time plausibility constraint such that $\phi_d(a, b) = 1$ if and only if b chronologically follows a. In this case, the date of a is the document's date of publication, which is given as metadata in HIPE-2020 and MHERCL, and the date of b is retrieved from Wikidata. We consider a wide set of time-related Wikidata properties reported in Table 16 in the Appendix. The properties' retrieval order has been set according to a qualitative analysis of the semantics of the properties' names and descriptions. The analysis aimed to ensure to retain, when available,

[11] Although CLOCQ is designed to output multiple candidate entities for each mention, for the sake of our evaluation we compare to the target only its top-ranked one.

[12] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, the model recommended by OpenAI as the optimal choice for cost-efficiency trade-off as of December 2024.

[13] https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/.

the most precise properties' values, such as P569 - DATE OF BIRTH (ranked first), against less precise ones, such as P585 - POINT IN TIME (ranked last).

5.2 Type constraint

The second criterion for candidate plausibility is comparing the types of two entities. Intuitively, if the mention is annotated as a human, entities of incompatible type e.g. building should be discarded.

We define $\phi_t \in \Phi$, the function that implements the type plausibility constraint such that $\phi_t(a, b) = 1$ if and only if the types of a and b are compatible. The types of a, i.e. the type of the mention, is the gold NER class assigned to the entity mention by the human annotators [for details about the types used in MHERCL, see Appendix C; for details about the types used in HIPE-2020, see Ehrmann et al. (2022b)]. The types of b are extracted by retrieving the value of the P31 - 'INSTANCE OF' Wikidata property and are manually mapped to NER classes used in MHERCL and HIPE-2020 by the same annotators that performed the annotation (see Sect. 3.2). Compatibility between NER classes used in MHERCL and HIPE-2020 and Wikidata types is further refined by compiling an *ad-hoc* taxonomy (see Appendix C) and checking whether the intersection between the types of a and b is not empty.

5.3 NIL heuristics

As reported in Sect. 4.2, we implement the possibility of predicting `NIL` in C-BLINK by assuming that if no candidate has a notably higher similarity with the mention than the others, a confident link is absent. In our experiments, we extensively test all `NIL` heuristics listed in Table 6 to assess their impact on the output. For threshold-based heuristics, we experiment with values $\tau \in [0, 1]$ with increments of 0.001. For machine learning approaches, we train classifiers on HIPE-2020 and evaluate their performance on MHERCL.

5.4 Results

This section reports the results of the experiments conducted with ELD, C-BLINK and the considered SotA models on HIPE-2020 and MHERCL. We report each constraint's influence and the heuristics' influence on the systems' accuracy. We evaluate our methods using the same evaluation metrics introduced by Ehrmann et al. (2022b). We follow the same experimental setting for all models: we consider a `NIL` annotation only if the model natively (or by extension, as in C-BLINK) supports `NIL` predictions. If the model does not provide any answer to our sentence, we consider it as wrongly linked. Table 7 reports the results for all the experiments on both MHERCL and HIPE-2020. Different observations emerge from Table 7, which are reported in the following paragraphs.

5.4.1 Off-the-shelf SotA entity linkers and LLMs fall short on HEL

Table 7 shows that SotA specialised EL models significantly underperform in historical document contexts, frequently failing to produce linkable entities for most sentences in both datasets. In particular, REL, CLOCQ, ExtEnD and ReLiK do not produce any linking candidate for mentions in the majority of the cases, as reported in Table 9. This issue underscores

a critical limitation in current approaches when faced with the specific challenges of historical texts, such as OCR errors and differences in morphosyntactic and semantic features between contemporary and historical language.

OCR errors significantly impact the performance of EL models. Table 8 presents the proportion of mentions affected by OCR errors over correctly and incorrectly linked mentions. OCR errors are more frequent in mentions linked incorrectly than in mentions linked correctly for all tested models. These errors are uncommon in typical NLP tasks and infrequent in pre-training corpora, leading PLMs to produce suboptimal representations. However, they are frequently found in historical documents (Table 3), making datasets like HIPE-2020 and MHERCL valuable for evaluating NLP robustness to OCR errors. As observed in Table 8, the proportions of mentions with OCR errors between correctly and incorrectly linked mentions are less unequal for the two LLMs tested (LLAMA 3.3 70B and GPT-4-o1-mini) compared to the other models. This suggests that recent LLMs with an high number of parameters are more robust to OCR errors than models specifically designed for the EL task, and this factor may account for their superior performance compared to all other SotA specialised models tested.

Moreover, specialised models jointly performing mentions recognition and linking generally underperform compared to the others. For instance, ReLiK and ExtEnD, which approach the EL task as an extraction problem, yield worse results than models requiring the gold NER information to be provided explicitly in their input, such as GENRE, mGENRE and BLINK. This underlines that historical datasets are also a valuable asset for NER. In particular, the high number of unique named entities in MHERCL makes it a convenient benchmark for assessing NER accuracy on long-tail entities. Indeed, we highlight that ReFinEd employs a similar architecture but achieves competitive results comparable to models explicitly receiving the gold NER information as input. This is because it is designed to utilize a broad spectrum of symbolic KB information that enhances its accuracy with long-tail and ambiguous entities.

GENRE and mGENRE require a prior named entity recognition phase and generally perform better than extraction-based models. However, they still underperform when compared to retrieval-based models such as BLINK. This trend is also supported by ELD's results, which operates similarly to a retrieval-based model. Despite their higher efficiency, generative-based models are unable to consider a large pool of linkable candidates since they must only consider a small subset of possible entities to maintain a tractable linking procedure.

Given these considerations, we argue that retrieval-based methods are less prone to popularity bias, as their encoding and retrieval phases can be designed to account for it. For instance, the performance of ELD relative to the other SotA specialised models tested indicates that it is less sensitive to popularity bias. Typically, SotA specialised models fine-tune the PLM on datasets suffering from popularity bias, as discussed in Sect. 3.5. If the PLM natively suffers from sub-optimal representations on long-tail entities, the fine-tuning approach will reinforce this aspect, further degrading its performances on historical documents. Since ELD relies on a PLM that is not specifically fine-tuned to accommodate the EL retrieval phase, it can rely on less biased entity representations.

The LLMs tested (LLAMA 3.3 70B and GPT-4 o1-mini) generally outperform the SotA specialised models in terms of accuracy. Their extensive computational resources and large-scale training data enable robustness to OCR noise and effective contextualization of sentence-level information based on the data encountered during training. Specifically,

Table 7 Baselines, ELD and C-BLINK results on HIPE2020 and MHERCL

Model	Φ	NIL heuristic	τ	HIPE-2020	MHERCL
Always NIL	–	–	–	0.43	0.31
van Hulst et al. (2020, REL)	–	–	–	0.16	0.10
De Cao et al. (2021, GENRE)	–	–	–	0.30	0.41
De Cao et al. (2022, mGENRE)	–	–	–	0.31	0.47
Christmann et al. (2022, CLOCQ*)	–	–	–	0.10	0.16
Ayoola et al. (2022, ReFinED*)	–	NIL in candidates	–	0.46	0.52
Barba et al. (2022, ExtEnD*)	–	–	–	0.04	0.25
Orlando et al. (2024, ReLiK*)	–	–	–	0.07	0.31
Wu et al. (2020, BLINK)	–	–	–	0.33	0.53
Wu et al. (2020, BLINK†)	–	–	–	0.34	0.52
LLAMA 3.3 70B	–	NIL allowed by prompt	–	0.60	0.61
GPT-4 o1-mini	–	NIL allowed by prompt	–	0.68	0.60
ELD	–	NIL in candidates	–	0.62	0.56
ELD	ϕ_d	NIL in candidates	–	0.63	0.57
ELD	ϕ_t	NIL in candidates	–	0.62	0.57
ELD	$\{\phi_d, \phi_t\}$	NIL in candidates	–	0.62	0.58
ELD†	–	NIL in candidates	–	0.46	0.49
ELD†	ϕ_d	NIL in candidates	–	0.47	0.45
ELD†	ϕ_t	NIL in candidates	–	0.48	0.49
ELD†	$\{\phi_d, \phi_t\}$	NIL in candidates	–	0.49	0.49
C-BLINK	ϕ_d	–	–	0.34	0.54
+	ϕ_t	–	–	0.34	0.54
+	$\{\phi_d, \phi_t\}$	–	–	0.35	0.54
+	$\{\phi_d, \phi_t\}$	Fixed threshold	0.443	0.66	**0.72**
+	$\{\phi_d, \phi_t\}$	Deviation from top	0.003	0.60	0.63
+	$\{\phi_d, \phi_t\}$	Deviation from median	0.003	0.60	0.63
+	$\{\phi_d, \phi_t\}$	Deviation from mean	0.000	0.57	0.59

Table 7 (continued)

Model	Φ	NIL heuristic	τ	HIPE-2020	MHERCL
+	$\{\phi_d, \phi_t\}$	Levenshtein	0.788	0.60	0.49
+	$\{\phi_d, \phi_t\}$	Jaccard	0.613	0.57	0.53
+	$\{\phi_d, \phi_t\}$	Hamming	0.402	0.58	0.48
+	$\{\phi_d, \phi_t\}$	SVM	–	0.65	0.69
+	$\{\phi_d, \phi_t\}$	Logistic regression	–	0.64	0.72
+	$\{\phi_d, \phi_t\}$	Decision Tree	–	**0.77**	0.59
C-BLINK†	ϕ_d	–	–	0.34	0.53
+	ϕ_t	–	–	0.34	0.53
+	$\{\phi_d, \phi_t\}$	–	–	0.34	0.54
+	$\{\phi_d, \phi_t\}$	Fixed threshold	0.626	0.54	0.49
+	$\{\phi_d, \phi_t\}$	Deviation from top	0.011	0.64	0.72
+	$\{\phi_d, \phi_t\}$	Deviation from median	0.025	0.67	0.74
+	$\{\phi_d, \phi_t\}$	Deviation from mean	0.022	0.67	**0.76**
+	$\{\phi_d, \phi_t\}$	Levenshtein	0.733	0.56	0.50
+	$\{\phi_d, \phi_t\}$	Jaccard	0.613	0.56	0.53
+	$\{\phi_d, \phi_t\}$	Hamming	0.467	0.57	0.47
+	$\{\phi_d, \phi_t\}$	SVM	–	0.66	0.71
+	$\{\phi_d, \phi_t\}$	Logistic Regression	–	0.65	0.74
+	$\{\phi_d, \phi_t\}$	Decision Tree	–	**0.76**	0.63

The highest F1 score is in bold. The application of which constraint (see Sects. 5.1 and 5.2), and NIL heuristics (see Table 6) is reported alongside their threshold τ, if needed. *AlwaysNIL* baseline is a dummy model that always predicts NIL for each mention. BLINK† and C-BLINK† do not rely on the re-ranking phase. ELD† does not use the replicator dynamics but selects the entity that is most similar to the mention. Models with * do not receive the gold named entity mention in the input

Table 8 Percentage of mentions presenting OCR errors over the total of correctly and incorrectly linked entities per model

Model	Dataset	OCR errors (%)	
		Over correct	Over wrong
GENRE	HIPE-2020	0.11	0.33
	MHERCL	0.05	0.19
mGENRE	HIPE-2020	0.09	0.34
	MHERCL	0.07	0.19
ReFinED	HIPE-2020	0.21	0.31
	MHERCL	0.07	0.20
LLAMA 3.3 70B	HIPE-2020	0.24	0.30
	MHERCL	0.11	0.17
GPT-4 o1-mini	HIPE-2020	0.26	0.28
	MHERCL	0.11	0.17
BLINK†	HIPE-2020	0.09	0.36
	MHERCL	0.08	0.18
BLINK	HIPE-2020	0.10	0.35
	MHERCL	0.09	0.19
C-BLINK⋆	HIPE-2020	0.19	0.43
	MHERCL	0.10	0.22
ELD⋆	HIPE-2020	0.20	0.39
	MHERCL	0.09	0.20

Models with a high number of empty answers, as reported in Table 9, are excluded

Table 9 Percentages of no responses (no-candidates output) for the models that frequently fail to provide any linkable candidates for the mentions, leading to invalid linking and poor performance

Model	Dataset	%
REL	HIPE-2020	0.70
	MHERCL	0.84
CLOCQ	HIPE-2020	0.80
	MHERCL	0.71
ExtEnD	HIPE-2020	0.95
	MHERCL	0.66
ReLiK	HIPE-2020	0.93
	MHERCL	0.64

Models from Table 7 for which such a scenario does not occur are excluded

LLAMA 3.3 70B, which provides access to the model but does not disclose its training dataset, achieves superior performance compared to all specialised SotA EL models tested on MHERCL. Although the computational effort required for this model is substantially higher than that of the other baselines, its high accuracy indicates that a distillation-based approach (Gu et al. 2024) could enhance retrieval-based systems by improving their retrieval components. Conversely, GPT-4 o1-mini demonstrates superior performance across all models on HIPE-2020 and achieves results comparable to LLAMA 3.3 70B on MHERCL. However, the lack of openly available model weights for GPT-4 limits its flexibility for integration with other approaches. Furthermore, caution is required regarding potential data contamination issues, particularly for the HIPE-2020 dataset, which has been available since 2020. This

concern aligns with observations raised in recent studies by Jacovi et al. (2023), Sainz et al. (2023), and Li and Flanigan (2024).

Table 10 quantitatively examines how entity popularity affects EL model performance, reporting the frequency with which more popular QIDs are chosen over less popular ones and the Spearman correlation between prediction correctness and entity popularity. C-BLINK⋆, ELD†, and ReFinED exhibit a lower tendency to favour popular QIDs on both MHERCL and HIPE-2020 datasets. For C-BLINK⋆ and ELD†, this reduced bias stems from the filtering approach discussed in Sects. 5.1 and 5.2, which excludes implausible entities based on temporal or type-related constraints, minimizing the chances of incorrectly selecting a more popular entity when it is logically implausible. ReFinED, designed specifically for robust performance with long-tail entities, demonstrates the lowest preference for more popular QIDs, as reflected in its Spearman correlation, which is remarkably lower than other models across both datasets. Conversely, the generative-based models, GENRE and

Table 10 Influence of entity popularity on EL performance

Model	Dataset	More popular QID chosen (%)	Spearman correlation
GENRE	HIPE-2020	**0.25**	**0.52***
	MHERCL	**0.16**	**0.55***
mGENRE	HIPE-2020	0.18	0.40*
	MHERCL	0.14	0.49*
ReFinED	HIPE-2020	<u>0.07</u>	0.17
	MHERCL	<u>0.05</u>	0.28*
BLINK†	HIPE-2020	0.12	0.40*
	MHERCL	0.11	0.31*
BLINK	HIPE-2020	0.14	0.40*
	MHERCL	0.10	0.33*
LLAMA 3.3 70B	HIPE-2020	0.10	0.32*
	MHERCL	0.05	0.21*
GPT-4 o1-mini	HIPE-2020	0.08	0.31*
	MHERCL	0.06	0.39*
C-BLINK$^{\dagger}_{\{\phi_d,\phi_t\}}$	HIPE-2020	0.14	0.40*
	MHERCL	0.08	0.34*
C-BLINK$_{\{\phi_d,\phi_t\}}$	HIPE-2020	0.14	0.38*
	MHERCL	0.10	0.31*
C-BLINK⋆	HIPE-2020	0.09	0.28*
	MHERCL	<u>0.03</u>	0.22*
ELD†	HIPE-2020	<u>0.07</u>	<u>−0.24</u>*
	MHERCL	0.06	0.18*
ELD$^{\dagger}_{\{\phi_d,\phi_t\}}$	HIPE-2020	<u>0.07</u>	−0.28*
	MHERCL	0.06	<u>0.05</u>
ELD	HIPE-2020	0.18	0.41*
	MHERCL	0.16	0.48*
ELD⋆	HIPE-2020	0.16	0.40*
	MHERCL	0.10	0.48*

For ELD and C-BLINK, we consider their best runs as indicated in Table 7. Spearman correlations marked with * denote statistical significance ($p < 0.01$) between popularity and correctness of the prediction. Models with a high number of no responses, as reported in Table 9, are excluded from this analysis. The lowest % increase and correlation coefficient are underlined, while the highest are marked in bold

mGENRE, show a stronger tendency to favour popular entities, reflecting a greater bias toward high-frequency entities.

In contrast, ELD's unsupervised approach relies on a sentence representation which is agnostic to the final task, eventually ensuring a fairer comparison between entities, which results in higher performance in HEL. Indeed, the results of ELD†, which does not use replicator dynamics, have a low correlation with the entities' popularity. Note, however, that using ELD with the replicator dynamics results in better quantitative performances (Table 7) but displays a positive correlation with popularity bias. This depends on the PLM employed during the encoding phase. If the PLM produces representations that "mildly" favour popular entities, the replicator dynamics process is akin to an online fine-tuning process, reinforcing the bias. It follows that there is a trade-off between quantitative performances and popularity bias using this approach. Further attention must be devoted to producing entity representations that reduce or eliminate bias towards popular entities.

The LLMs tested (LLAMA 3.3 70B and GPT-4 o1-mini) appear more robust to popularity variations, as evidenced by their percentage of more popular QIDs chosen over less popular ones and Spearman correlations comparable to those of C-BLINK⋆, ELD†, which are designed to logically exclude implausible entities, and ReFinED, which is designed to achieve better performance on long-tail knowledge scenarios. Their massive-scale training data and architectural complexity make them competitive in linking less popular entities.

Interestingly, all the specialised SotA EL models perform better on MHERCL than on HIPE-2020, despite MHERCL being composed of less popular entities as compared to HIPE-2020 (see Sect. 3.5). This discrepancy is evident in BLINK, which, while underperforming on HIPE-2020, emerges as the best-performing baseline on MHERCL. The difference in performance can be attributed to the higher incidence of NIL entities in HIPE-2020 compared to MHERCL. Employing a conservative strategy that invariably predicts NIL already establishes a robust baseline. In contrast, the lower prevalence of NIL entities in MHERCL facilitates more effective testing of SotA entity linkers on historical documents, providing a clearer assessment of their capabilities.

5.4.2 Time and type constraints enable performance improvement on HEL

The results from Table 7 prove that plausibility filters (as introduced in Sects. 5.1 and 5.2) are a straightforward way of enhancing EL methods when historical documents are considered. Both ELD and C-BLINK demonstrate equal or superior performance when filters are enabled.

For instance, C-BLINK with time and type constraints can link historical entities where most the other models fail. In the first example of Table 12, all the tested models but C-BLINK⋆ and LLAMA 3.3 70B produce links to entities that are implausible because born after the date of issue of the historical periodical from which the sentence was extrapolated. Note that ReFinED, ELD⋆ and GPT-4 o1-mini are overly conservative in predicting a NIL link. Similarly, in the second example of Table 12, all the tested models but C-BLINK⋆ and LLAMA 3.3 70B produce inconsistent links regarding the entity type. C-BLINK⋆ leverages the implemented plausibility filters to correctly predict the expected entities. Such a successful linking happens despite the OCR error in the mention, demonstrating that plausibility filters can also be helpful in cases where the mention presents OCR noise. Once again, ReFinED and ELD⋆ conservatively predict NIL. Interestingly, LLAMA 3.3 70B can overcome the

OCR errors affecting the mention and leverages the context of the sentence to correctly identify the expected mention, while GPT-4 o1-mini does not show the same behaviour and rather predicts NIL. Multiple reasons can explain this aspect, given the opaque inference mechanism that characterizes LLMs. We posit that the different instruction tuning processes employed in specializing both models result in a model, GPT-4 o1-mini, being more conservative in its predictions. This also explains the higher performances of this model in Table 7, where models that consistently rely on NIL links outperform the other models. In the third example, C-BLINK* is the only model able to predict the correct entity. This success is attributed to its plausibility filters, which restrict the model to considering only entities that are valid in terms of both type and time information. In contrast, the other models fail by either suggesting an entity with an incompatible type or one with an implausible date. In the final example, instead, ELD* is the only model that predicts the correct entity thanks to the application of the plausibility filters, while C-BLINK* is excessively conservative and predicts a NIL link. In this case, the other models do not find the correct entity, either because of type or date implausibility.

In Table 11, we report the accuracy and F1-score for each model in predicting a plausible entity by framing it as a classification task. A predicted entity is considered plausible with respect to the target entity if its time information, retrieved as discussed in Sect. 5.1, precedes the issue date of the source document containing the sentence where the target entity appears. Similarly, it is considered plausible if its type matches that of the target entity. BLINK, GENRE, mGENRE and LLAMA 3.3 70B obtain good performances on both time and type plausibility, even when compared to C-BLINK. This provides evidence that when relying on a large amount of training data, those models can infer some level of plausibil-

Table 11 Accuracy and F1 score computed on the type and time plausibility of SotA models and our proposed models' predictions

Model	Dataset	Year_A	Year_{F1}	Type_A	Type_{F1}
GENRE	HIPE-2020	0.67	0.80	0.60	0.75
	MHERCL	0.70	0.82	0.77	0.87
mGENRE	HIPE-2020	0.63	0.77	0.54	0.70
	MHERCL	0.76	0.87	0.77	0.87
ReFinED	HIPE-2020	0.27	0.43	0.25	0.40
	MHERCL	0.55	0.71	0.54	0.70
BLINK	HIPE-2020	0.74	0.85	0.66	0.80
	MHERCL	0.85	0.92	0.85	**0.92**
BLINK†	HIPE-2020	0.70	0.82	0.67	0.80
	MHERCL	0.79	0.88	0.80	0.89
LLAMA 3.3 70B	HIPE-2020	**0.79**	**0.89**	**0.68**	**0.81**
	MHERCL	0.80	0.89	0.74	0.85
GPT-4 o1-mini	HIPE-2020	0.72	0.84	0.60	0.75
	MHERCL	0.63	0.77	0.61	0.76
C-BLINK*	HIPE-2020	0.74	0.85	0.66	0.80
	MHERCL	**0.87**	**0.93**	**0.86**	**0.92**
ELD*	HIPE-2020	0.51	0.68	0.44	0.61
	MHERCL	0.51	0.68	0.50	0.66

Models with a high number of no responses, as reported in Table 9, are excluded from this analysis. Subscripts indicate the metric. An empty answer is considered wrong. Mentions labelled with NIL are excluded. The best results are represented in bold. The worst results are underlined

ity. In particular, LLAMA 3.3 70B outperforms all other models in plausibility accuracy on HIPE-2020, even the ones including filters. Nonetheless, its performances in Table 7 are significantly lower than the ones of GPT-4 o1-mini. This is explained by the tendency of GPT-4 o1-mini to favour NIL predictions, as seen in the examples of Table 12, and underscores the importance of this aspect in HEL. We highlight that the plausibility performances of LLAMA 3.3 70B are not consistent between HIPE-2020 and MHERCL. This further reinforces the argument that MHERCL is a valuable dataset in the NLP landscape.

Note that the application of time and type constraints on C-BLINK and ELD do not always restrict the pool of candidates to only those that are plausible. This is due to the mapping between NERC types and Wikidata types that we designed. Nonetheless, C-BLINK demonstrates a higher ability to propose a plausible candidate when compared to BLINK.

We remark that plausibility constraints are crucial for enabling the application of the NIL heuristics discussed in Sect. 5.3. These heuristics assume that all candidates considered by the model are plausible options. In cases where no plausible candidate has a significantly higher similarity score than the others, we assume that the model cannot confidently establish a reliable link and should instead output NIL. This heuristic would be impossible to apply without a plausibility assumption since implausible entities might still display a high similarity score due to the PLMs tendency to favour similarity with popular entities.

5.4.3 Effective models for HEL must take NIL into account

We experiment with several NIL heuristics and observe different outcomes. In general, machine learning-based heuristics fail to generalize to different distributions, as seen in C-BLINK relying on decision trees. The best approach is C-BLINK† with the deviation from the mean heuristics. By avoiding the re-ranking phase, the scores produced by BLINK† during the candidate scoring phase are more reliable in NIL prediction. Since we remove every implausible candidate, the model can conservatively avoid a wrong link if no remaining candidate significantly stands out. NIL heuristics based on the similarity between superficial mentions are less effective than expected. This can be attributed to OCR errors and the different language registries employed by historical documents compared to the one found in the reference KB.

Our initial hypothesis on the importance of managing NIL predictions is empirically confirmed by the fact that the best NIL heuristic for C-BLINK largely outperforms all the other models on both HIPE-2020 and MHERCL. Despite allowing tested LLMs to link named entity mentions to NIL when no suitable Wikipedia page title could be suggested, the proposed C-BLINK† with the deviation from the mean heuristic achieves the highest F1 score. The results highlight that a simpler retrieval-based neural entity linking model enhanced with targeted heuristics can outperform larger, more expensive, and closed-source models in long-tail, domain-specific scenarios requiring robust NIL handling.

Additionally, it is notable that ELD using replicator dynamics, with type- and time-plausibility candidates filtering logic activated and NIL in candidates outperforms LLAMA 3.3 70B on HIPE-2020. On MHERCL, it still performs competitively, falling short of LLAMA 3.3 70B's performance by only 0.03 F1 and GPT-4 o1-mini's by just 0.02 F1.

Table 13 provides an analysis of the incidence of NIL predictions for the best C-BLINK and ELD models in Table 7 on MHERCL and HIPE-2020. The analysis is divided between mentions that should be linked to NIL and the others. Furthermore, whether the linking target is among the list of candidates at the disposal of the model is analyzed.

Table 12 Examples of predictions made on MHERCL and HIPE-2020

The Quarterly Musical Magazine And Review, 1825 (MHERCL)				
M. Barriere (`Q726908, Jean-Baptiste Barrière, person, 1707`) has published four sets of quartets, and several Symphonies, concertos, trios, and duets				
Model	Prediction	Wikipedia title	Type	Year
BLINK	Q3086402	Françoise Barrière	**Human**	1944
GENRE	Q56537639	Barrière	Family name	
mGENRE	Q3588326	Émile Barrière	**Human**	1902
ReFinED	NIL			
LLAMA 3.3 70B	**Q726908**	**Jean-Baptiste Barrière**	**Human**	**1707**
GPT-4 o1-mini	NIL			
C-BLINK*	**Q726908**	**Jean-Baptiste Barrière**	**Human**	**1707**
ELD*	NIL			
The Musical Times, 1873 (MHERCL)				
On the 14th ult the South Wales Choral Union visited Marlborough House (`Q565532, Marlborough House, building, 1711`), *by express desire of His Royal Highness the Prince of Wales*				
Model	Prediction	Wikipedia title	Type	Year
BLINK	Q539528	Marlborough, Wiltshire	Market town	
GENRE	Q539528	Marlborough, Wiltshire	Market town	
mGENRE	Q1264586	Duke of Marlborough (title)	Noble title	
ReFinED	N/A			
LLAMA 3.3 70B	**Q565532**	**Marlborough House**	**Mansion**	**1711**
GPT-4 o1-mini	NIL			
C-BLINK*	**Q565532**	**Marlborough House**	**Mansion**	**1711**
ELD*	NIL			
The Quarterly Musical Magazine and Review, 1828 (MHERCL)				
Furno (`Q555808, Giovanni Furno, person, 1748`), *who fills this post, is equally active and skilful*				
Model	Prediction	Wikipedia title	Type	Year
BLINK	Q920465	Carlo Furno	**Human**	1921
GENRE	Q1475080	Furno	Wikimedia Disambiguation Page	
mGENRE	Q1475080	Furno	Wikimedia Disambiguation Page	
ReFinED	Q104052585	Furno	**Human**	
LLAMA 3.3 70B	Q517279	Antonio Maria Vegliò	**Human**	1938
GPT-4 o1-mini	NIL			
C-BLINK*	**Q555808**	**Giovanni Furno**	**Human**	**1748**
ELD*	NIL			
sn83030483, 1790 (HIPE-2020)				
JOHN SULLIVAN. (`Q886420, John Sullivan, person, 1740`)				
Model	Prediction	Wikipedia title	Type	Year
BLINK	Q6259624	John Sullivan (VC)	**Human**	1830
GENRE	Q6259628	John Sullivan (pitcher)	**Human**	1894
mGENRE	Q11082552	John Sullivan	Wikimedia human name disambiguation page	

Table 12 (continued)

sn83030483, 1790 (HIPE-2020)				
JOHN SULLIVAN. (*Q886420, John Sullivan, person, 1740*)				
Model	Prediction	Wikipedia title	Type	Year
ReFinED	Q3396287	John Sullivan (writer)	**Human**	1946
LLAMA 3.3 70B	Q11082552	John Sullivan	Wikimedia human name disambiguation page	
GPT-4 o1-mini	Q11082552	John Sullivan	Wikimedia human name disambiguation page	
C-BLINK⋆	NIL			
ELD⋆	**Q886420**	**John Sullivan (general)**	**Human**	**1740**

The target mention is underlined, and its QID, Wikipedia page title, NEC class and Wikidata year are in brackets. We report the predicted QID, Wikipedia page title, and Wikidata type and time information. Matching information is in bold. C-BLINK⋆ and ELD⋆ are the best models in Table 7. Models with a high number of no responses (Table 9) are excluded

Table 13 Error analysis on MHERCL for the best runs of C-BLINK and ELD

Model	Target	Is target in top-k candidates?	Wrong prediction	Errors (%)	
				HIPE-2020	MHERCL
C-BLINK⋆	NIL	–	QID	0.14	0.17
	QID	×	NIL	**0.47**	0.31
		×	QID	0.05	0.06
		✓	NIL	0.26	**0.38**
		✓	QID	0.07	0.08
ELD⋆	NIL	×	QID	0.09	0.07
		✓	QID	0.05	0.06
	QID	×	NIL	**0.64**	**0.64**
		×	QID	0.11	0.18
		✓	NIL	0.0	0.002
		✓	QID	0.03	0.03

The *Target* is the gold annotation, which can be NIL or a QID. The target can or cannot be among the candidates retrieved by the models. Note that NIL can never be among C-BLINK candidates, but it can be retrieved by ELD. The highest error type percentages per each model are indicated in bold text, while the lower ones are underlined

For C-BLINK, the most frequent error on MHERCL occurs when the model predicts NIL despite the target QID being among the candidates. This is a consequence of our NIL prediction heuristics. Even though this approach allows better performances, they rely on the scores produced by the model and assume that their magnitude is a reliable proxy for their fitness as a final link. However, the retriever (and re-ranker) models are not explicitly trained for this task. It might hence happen that the target candidate is not perceived as significantly different from the other plausible results, hence resulting in the inference of NIL. Integrating a plausibility mechanism that can propagate the decision to the encoder is a promising approach that could result in more robust entity encoders that produce more reliable similarity scores. The model's second most frequent error on MHERCL and first most frequent error on HIPE-2020 happens when the model

predicts NIL because the target QID is not within the candidates considered. Such a behaviour can also be observed in ELD, where the highest % of error types relate to cases where the target QID is not among the linking candidates, and the model predicts NIL. The least frequent errors for both models include cases where the target QID is absent, but the model predicts a wrong QID, or where the target QID is present, and the model selects a different one. In particular, ELD rarely prefers NIL over a correct QID and chooses a different (wrong) QID only in 3% of the cases. As predicting NIL is optimal when the target QID is not in the candidate set, such a conservative approach is beneficial in scenarios in which precise prediction of QIDs is key, as in long-tail, domain-specific knowledge extraction scenarios. In such contexts, where low-popularity named entities are prevalent, it is often preferable to avoid incorrect linkages by opting for NIL rather than risking an inaccurate association. Thus, while the model's conservatism may lead to a higher rate of NIL predictions, it ultimately supports more precise outcomes in cases where the entities in question are less popular.

6 Conclusion

In this paper, we introduced MHERCL, a gold standard benchmark of annotated English sentences for historical NER, NEC and EL. This dataset comprises sentences from music historical periodicals published between 1823 and 1900 and collected within the Polifonia Corpus[1]. It is intended to support further study in HEL and, more generally, on the recognition and linking of long-tail and domain-specific named entities.

We also presented and evaluated ELD, an unsupervised model for EL, and C-BLINK, an extended version of BLINK incorporating plausibility filters that leverage knowledge from trusted KGs, in our case Wikidata. ELD demonstrates that an unsupervised model can be effective in HEL, addressing the popularity bias in training datasets by eliminating the need for a supervised training phase. Implementing ELD using replicator dynamics from game theory outperforms all baselines. The implementation of time and type constraints within C-BLINK has enabled a significant improvement of the model's performance on the tested historical documents benchmarks. Assuming that all candidates are plausible based on time and type allows C-BLINK to integrate heuristics predicting entities that are not in the KB of reference (NIL links), outperforming all the other tested models. When plausibility filters are applied, ELD's performance improves further, underscoring the value of integrating these constraints in HEL tasks.

Future work will focus on expanding the benchmark to include annotated sentences in other languages, thereby addressing the gap in resource availability for NER, NEC, and EL tasks, as well as the QA task (Graciotti et al., 2024). Moreover, we aim to test OCR post-correction methods (Ramirez-Orta et al., 2022) to mitigate the impact of OCR errors on candidate generation and linking. Additionally, we plan to complement OCR post-correction with data augmentation to enhance lexical variability (Lacerra et al., 2021a, b). Additionally, we plan to explore methods for integrating plausibility constraints into the training process of EL models and evaluate these approaches on standard EL benchmarks. Finally, the competitive accuracy of LLMs when prompted in a zero-shot fashion in the context of HEL is a promising approach, proving that their huge pre-training process helps in coupling better with the specifics of historical documents (OCR errors, morphosyntactic differences from contemporary language, etc.) as well as popularity bias. In particular, the filtered retrieval approach employed for C-BLINK

can be integrated into Retrieval Augmented Generation (RAG) methods, delegating to LLMs the final decision on the entity that is better suited as a link, exploiting their resilience to OCR errors. Moreover, a particularly promising approach is to integrate an open-weights LLM, such as LLAMA 3.3 70B, with the approach used by (m)GENRE. Specifically, restricting the generation of candidates to those considered plausible would allow these models to enforce time and type plausibility, enhancing their capabilities and potentially improving their sensitivity to `NIL` predictions.

Appendix A: annotation guidelines

General guidelines
The annotators were instructed to consider a named entity a real-world thing indicating a unique individual through a proper noun (Jurafsky and Martin 2023). The annotators were introduced to a typical NLP pipeline (tokenization, part-of-speech tagging, word sense disambiguation, NER, NEC and EL), emphasising the NER, NEC and EL tasks.
Nested entities
The annotators were instructed not to annotate nested named entities. In fact, in cases of named entities such as *Account of the Oxford Commemoration Concerts*, the only valid named entity would be the main one (in our example, the entity of type PUBLICATION *Account of the Oxford Commemoration Concerts*). Therefore, in our example, "Oxford" would not be annotated as an entity of type CITY.
Noisy entities
The annotator is requested to record whether the named entity's superficial mention presents OCR errors. This strategy has allowed us to keep track of how many named entities' superficial mentions were *noisy*, namely impacted by OCR errors (see Sect. 3.4).
Named entity classification
The named entities' types were assigned according to the AMR annotation guidance instructions and selected from the AMR named entity types list.[14] Sticking to the AMR annotation guidance instructions, the annotators were instructed to select the most specific type possible.
Entity linking
The annotators were instructed to link the named entities recognized to their corresponding QID.
`NIL` *links*
The annotators were instructed to assign the string `NIL` to named entities not present in Wikidata.

Appendix B: MHERCL format

MHERCL is released in CoNLL-U format[15] through its dedicated GitHub repository[16] and in JSONL format in HuggingFace[1].

[14] Available at the URL https://github.com/amrisi/amr-guidelines/blob/master/amr.md#named-entities, last time accessed on August 12th 2024.

[15] https://universaldependencies.org/format.html

[16] Available at the URL https://github.com/polifonia-project/historical-entity-linking/blob/main/benchmark/v1.0/mhercl_v1.0.tsv.

MHERCL CoNLL-U format release complies with HIPE-2022 data (based on IOB and CoNLL-U), facilitating future integration. An example of an annotated sentence of MHERCL in CoNLL-U format is reported below:

```
#document_id:DwightSJournalOfMusic__1873-016.txt_762
#document_date:1873
#sent_text:A native of Parma, ateighteen years of age Jong was
    received into the Conservatory of Music of that town, where
    Jong soon made himself.a name as the most promising pupil of
     the institution.
A       O       _
native  O       _
of      O       _
Parma   B-city  Q2683
,       O       _
ateighteen      O       _
years   O       _
of      O       _
age     O       _
Jong    B-person        NIL
was     O       _
received        O       _
into    O       _
the     O       _
Conservatory    B-school        Q1439627
of      I-school        Q1439627
Music   I-school        Q1439627
of      O       _
that    O       _
town    O       _
,       O       _
where   O       _
Jong    B-person        NIL
soon    O       _
made    O       _
himself.a       O       _
name    O       _
as      O       _
the     O       _
most    O       _
promising       O       _
pupil   O       _
of      O       _
the     O       _
institution     O       _
.       O       _
```

- In the above example, we can see that rows beginning with # contain metadata. In more detail:#DOCUMENT_ID, provides the identifier (ID) of the document the sentence is extrapolated from;
- #DOCUMENT_DATE, provides the date in which the document was published;

- #SENT_TEXT, provides the text of the annotated sentence.. The columns can be explained as follows:COLUMN 0 reports the sentence's tokens, one per row;
- COLUMN 1, reports IOB[17] tags and type label of the named entity;
- COLUMN 2, reports the QID assigned by the annotators to the identified named entity.

Appendix C: named entities classification

C.1: named entities types

The types assigned to named entities in MHERCL are selected from the AMR named entity types list.[18] In Table 14, we report all the entity types assigned to named entities in MHERCL.

[17] According to IOB notation tagging format, each token which constitutes the beginning of a named entity string is assigned the label 'B' (an abbreviation of 'beginning'). Each token within a named entity string is labelled 'I' (an abbreviation for 'inside'). Each token outside a named entity string is labelled 'O' (an abbreviation for 'outside').

[18] We refer to their list as documented in the official GitHub repository, accessible at the URL https://github.com/amrisi/amr-guidelines/blob/master/amr.md#named-entities.

Table 14 Named entity types occurring in MHERCL

Named entity type	# Occurrences	Named entity type	# Occurrences
Person	1246	Song	3
City	251	Concert	3
Music	188	Location	3
Organization	94	River	3
Work-of-art	85	Museum	3
Country	78	Newspaper	3
Building	52	Country-region	3
Opera	52	Symphony	2
Theatre	41	Religious-group	2
Worship-place	41	Thing	2
Publication	26	Person	2
Book	24	Family	2
Road	24	Language	1
Company	16	Band	1
School	16	Province	1
City-district	13	Island	1
Magazine	9	Park	1
Event	8	Empire	1
Festival	8	Hotel	1
Street	7	Scholarship	1
Mountain	6	Institution	1
University	6	Village	1
Government-organization	5	Town	1
College	4	Books	1
Facility	4	Person (fictional character)	1
Local-region	4	Lake	1
County	4	Hall	1
Continent	4	Society	1
Journal	3	Military	1
Square	3		

C.2: named entities types taxonomy

As described in Sect. 5.2 of this paper, the named entity types are used in C-BLINK and ELD functionalities that assess the plausibility of the candidates according to their types. To optimize such a filter, we developed a taxonomy categorizing each named entity type occurring in MHERCL into a hierarchy of types. When applying the type filter functionality, we expanded the dataset's gold type for each entity by including their more generic types from this taxonomy. The taxonomy is reported in Table 15.

Table 15 Complete taxonomy used for expanding named entities' types to apply type constraints

Type	Sub-types
B-event	B-concert, B-festival
B-facility	B-street, B-road, B-park
B-building	B-theatre, B-university, B-worship-place, B-museum, B-college, B-company, B-school, B-hall
B-language	
B-organization	B-theatre, B-university, B-worship-place, B-museum, B-college, B-company, B-school, B-empire, B-government-organization, B-religious-group, B-band
B-person	
B-publication	B-book, B-magazine, B-newspaper, B-journal
B-work-of-art	B-music, B-opera, B-symphony, B-book, B-song
B-location	B-park, B-hall, B-city, B-city-district, B-continent, B-country, B-county, B-local-region, B-mountain, B-road, B-square, B-country-region, B-province, B-island

Appendix D: time-related Wikidata properties retrieval

As described in Sect. 5.1 of this paper, time-related Wikidata properties are used in C-BLINK and ELD functionalities that assess the plausibility of the linking candidates according to their chronological consistency (time plausibility).

We report the time-related Wikidata properties that we retrieve in Table 16. The retrieval order is based on a qualitative analysis of the properties' names and descriptions to prioritize those values that can more precisely determine the plausibility of an entity, such as P569 - DATE OF BIRTH (ranked first), over more generic ones, like P585 - POINT IN TIME (ranked last).

Table 16 Time-related Wikidata properties used for candidate retention functionalities

Retrieval order	Wikidata property	Name	Description
1	P569	DATE OF BIRTH	Date on which the subject was born
2	P571	INCEPTION	Time when an entity begins to exist
3	P1619	DATE OF OFFICIAL OPENING	Date or point in time an event, museum, theater etc. officially opened
4	P1191	DATE OF FIRST PERFORMANCE	Date a work was first debuted, performed or live-broadcasted
5	P10135	RECORDING DATE	The date when a recording was made
6	P577	PUBLICATION DATE	Date or point in time when a work was first published or released
7	P575	TIME OF DISCOVERY OR INVENTION	Date or point in time when the item was discovered or invented
8	P1317	FLORUIT	Date when the person was known to be active or alive, when birth or death not documented
9	P7124	DATE OF THE FIRST ONE	Qualifier: when the first element of a quantity appeared/took place
10	P10673	DEBUT DATE	Date when a person or group is considered to have "debuted"
11	P9448	INTRODUCED ON	Date when a bill was introduced into a legislature
12	P6949	ANNOUNCEMENT DATE	Time of the first public presentation of a subject by the creator, of information by the media
13	P729	SERVICE ENTRY	Date or point in time on which a piece or class of equipment entered operational service
14	P2031	WORK PERIOD (START)	Start of period during which a person or group flourished (fl. = "floruit") in their professional activity
15	P585	POINT IN TIME	Time and date something took place, existed or a statement was true

Appendix E: prompt for LLMs evaluation on HEL

This appendix provides the prompt used to test LLMs (LLAMA 3.3 70B and GPT-4 o1-mini) in the experimental setting reported in Sect. 5. The prompt, designed to query the model for the Wikipedia page name corresponding to a specific entity mention within a given sentence, is reported below:

```
In the sentence: <sentence> what is the Wikipedia page name of the entity <mention>? Answer with the page title only. If no Wikipedia page is appropriate, answer "NIL". Do not write anything else.
```

Acknowledgements This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 101004746. The authors acknowledge Dr Enrico Daga's feedback on this manuscript, which provided an insightful external perspective.

Author contributions A. G. contributed to the paper conceptualisation and in writing its original draft, curated the data collection for the MHERCL benchmark, curated the design of the annotation guidelines and coordinated the annotation campaign, curated the knowledge-based resources required to develop C-BLINK and contributed to the design of time and type constraints, contributed to the software development, evaluation and testing, and the analysis of the experimental results. N. L. curated the methodology design for linking entities absent in the Knowledge Base of reference (NIL links), curated the methodology design, the software development and evaluation of C-BLINK and contributed to the design of time and type constraints, and contributed to the analysis of the experimental results. R. T. acted as the scientific coordinator of the paper, curated the paper conceptualisation and the writing of its original draft, curated the design of the methodology, the software development and evaluation of Entity Linking Dynamics, and contributed to the analysis of the experimental results. V. P. acted as the supervisor of the work and scientific coordinator of the paper, secured the funding and administrated the resources. All the authors contributed to the writing, reviewing, and editing of the paper.

Funding Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement.
This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 101004746. Nicolas Lazzari is supported by the FAIR - Future Artificial Intelligence Research Foundation as part of the Grant Agreement MUR n. 341, code PE00000013 CUP 53C22003630006.

Data availability All the code for our experiments is publicly available on GitHub (Available at the URL https://github.com/polifonia-project/historical-entity-linking.). The MHERCL dataset is available in the same GitHub repository and on HuggingFace (Available at the URL https://huggingface.co/datasets/n28div/MHERCL.).

Declarations

Conflict of interest The authors have no conflict of interest to declare that are relevant to the content of this article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Agarwal P, Strötgen J, Corro L, Hoffart J, Weikum G (2018) diaNED: time-aware named entity disambiguation for diachronic corpora. In: Gurevych I, Miyao Y (eds) Proceedings of the 56th annual meeting of the Association for Computational Linguistics (volume 2: short papers). Association for Computational Linguistics, Melbourne, Australia, pp 686–693. https://doi.org/10.18653/v1/P18-2109

Alani H, Guo Z, Barbosa D (2018) Robust named entity disambiguation with random walks. Semant Web J 9(4):459–479. https://doi.org/10.3233/SW-170273

Ayoola T, Tyagi S, Fisher J, Christodoulopoulos C, Pierleoni A (2022) ReFinED: an efficient zero-shot-capable approach to end-to-end entity linking. In: Loukina A, Gangadharaiah R, Min B (eds) Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies: industry track. Association for Computational Linguistics, Hybrid: Seattle, United States and Online, pp 209–220. https://doi.org/10.18653/v1/2022.naacl-industry.24

Banarescu L, Bonial C, Cai S, Georgescu M, Griffitt K, Hermjakob U, Knight K, Koehn P, Palmer M, Schneider N (2013) Abstract Meaning Representation for sembanking. In: Pareja-Lora A, Liakata M, Dipper S (eds) Proceedings of the 7th linguistic annotation workshop and interoperability with discourse. Association for Computational Linguistics, Sofia, Bulgaria, pp 178–186. https://aclanthology.org/W13-2322

Barba E, Procopio L, Navigli R (2022) ExtEnD: extractive entity disambiguation. In: Muresan S, Nakov P, Villavicencio A (eds) Proceedings of the 60th annual meeting of the Association for Computational Linguistics (volume 1: long papers). Association for Computational Linguistics, Dublin, Ireland, pp 2478–2488. https://doi.org/10.18653/v1/2022.acl-long.177

Benkhedda Y, Skapars A, Schlegel V, Nenadic G, Batista-Navarro R (2024) Enriching the metadata of community-generated digital content through entity linking: an evaluative comparison of state-of-the-art models. In: Bizzoni Y, Degaetano-Ortlieb S, Kazantseva A, Szpakowicz S (eds) Proceedings of the 8th joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature (LaTeCH-CLfL 2024). Association for Computational Linguistics, St. Julians, Malta, pp 213–220. https://aclanthology.org/2024.latechclfl-1.20

Blouin B, Armand C, Henriot C (2024) A dataset for named entity recognition and entity linking in Chinese historical newspapers. In: Calzolari N, Kan M-Y, Hoste V, Lenci A, Sakti S, Xue N (eds) Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024). ELRA and ICCL, Torino, Italia, pp 385–394. https://aclanthology.org/2024.lrec-main.35

Boros E, Pontes EL, Cabrera-Diego LA, Hamdi A, Moreno JG, Sidère N, Doucet A (2020) Robust named entity recognition and linking on historical multilingual documents. In: Working notes of CLEF 2020, vol 2696. pp 1–17. https://ceur-ws.org/Vol-2696/paper_171.pdf

Boros E, González-Gallardo C-E, Giamphy E, Hamdi A, Moreno JG, Doucet A (2022) Knowledge-based contexts for historical named entity recognition and linking. In: Faggioli G, Ferro N, Hanbury A, Potthast M (eds) Proceedings of the working notes of CLEF 2022—conference and labs of the evaluation forum. CEUR workshop proceedings, vol 3180. CEUR-WS.org, Bologna, Italy, pp 1064–1078. http://ceur-ws.org/Vol-3180/paper-84.pdf

Bunescu R, Pașca M (2006) Using encyclopedic knowledge for named entity disambiguation. In: McCarthy D, Wintner S (eds) 11th conference of the European chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Trento, Italy, pp 9–16. https://aclanthology.org/E06-1002

Castro S (2017) Fast Krippendorff: fast computation of Krippendorff's alpha agreement measure. GitHub. https://github.com/pln-fing-udelar/fast-krippendorff. Accessed 13 Aug 2024

Chen A, Gudipati P, Longpre S, Ling X, Singh S (2021) Evaluating entity disambiguation and the role of popularity in retrieval-based NLP. 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). Association for Computational Linguistics, Online, pp 4472–4485. https://doi.org/10.18653/v1/2021.acl-long.345

Christmann P, Saha Roy R, Weikum G (2022) Beyond NED: fast and effective search space reduction for complex question answering over knowledge bases. In: Proceedings of the fifteenth ACM international conference on web search and data mining. WSDM '22. Association for Computing Machinery, New York, NY, USA, pp 172–180. https://doi.org/10.1145/3488560.3498488

De Cao N, Izacard G, Riedel S, Petroni F (2021) Autoregressive entity retrieval. In: 9th international conference on learning representations. OpenReview.net, Online, Austria. https://openreview.net/forum?id=5k8F6UU39V

De Cao N, Wu L, Popat K, Artetxe M, Goyal N, Plekhanov M, Zettlemoyer L, Cancedda N, Riedel S, Petroni F (2022) Multilingual autoregressive entity linking. Trans Assoc Comput Linguist 10:274–290. https://doi.org/10.1162/tacl_a_00460

Devlin J, Chang M-W, Lee K, Toutanova K (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). Association for Computational Linguistics, Minneapolis, Minnesota, pp 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dong H, Chen J, He Y, Liu Y, Horrocks I (2023) Reveal the unknown: out-of-knowledge-base mention discovery with entity linking. In: Proceedings of the 32nd ACM international conference on information and knowledge management. CIKM '23. Association for Computing Machinery, New York, NY, USA, pp 452–462. https://doi.org/10.1145/3583780.3615036

Ehrmann M, Colavizza G, Rochat Y, Kaplan F (2016) Diachronic evaluation of NER systems on old newspapers. In: Dipper S, Neubarth F, Zinsmeister H (eds) Proceedings of the 13th conference on natural language processing (KONVENS 2016). Bochumer Linguistische Arbeitsberichte, Bochum, Germany, pp 97–107. https://infoscience.epfl.ch/record/221391?v=pdf

Ehrmann M, Romanello M, Fluckiger A, Clematide S (2020a) Extended overview of CLEF HIPE 2020: named entity processing on historical newspapers. In: Working notes of CLEF 2020—conference and labs of the evaluation forum, vol 2696. Thessaloniki, Greece, p 38. https://doi.org/10.5281/zenodo.4117566

Ehrmann M, Romanello M, Flückiger A, Clematide S (2020b) Overview of CLEF HIPE 2020: named entity recognition and linking on historical newspapers. In: Experimental IR meets multilinguality, multimodality, and interaction: 11th international conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, proceedings. Springer, Berlin, Heidelberg, pp 288–310. https://doi.org/10.1007/978-3-030-58219-7_21

Ehrmann M, Romanello M, Doucet A, Clematide S (2022a) Introducing the HIPE 2022 shared task: named entity recognition and linking in multilingual historical documents. In: Hagen M, Verberne S, Macdonald C, Seifert C, Balog K, Nørvåg K, Setty V (eds) Advances in information retrieval, vol 13186. Springer International Publishing, Cham, pp 347–354. https://doi.org/10.1007/978-3-030-99739-7_44

Ehrmann M, Romanello M, Najem-Meyer S, Doucet A, Clematide S (2022b) Extended overview of HIPE-2022: named entity recognition and linking in multilingual historical documents. In: Faggioli G, Ferro N, Hanbury A, Potthast M (eds) Working notes of CLEF 2022—conference and labs of the evaluation forum (CLEF), vol 3180. CEUR-WS, Aachen, pp 1038–1063. https://doi.org/10.5281/zenodo.6979577

Ehrmann M, Hamdi A, Pontes EL, Romanello M, Doucet A (2023) Named entity recognition and classification in historical documents: a survey. ACM Comput Surv. https://doi.org/10.1145/3604931

Gabrilovich E, Ringgaard M, Subramanya A (2013) FACC1: freebase annotation of ClueWeb corpora, version 1 (release date 2013-06-26, format version 1, correction level 0). Web Download. http://lemurproject.org/clueweb09/, http://lemurproject.org/clueweb12/

Graciotti A, Presutti V, Tripodi R (2024) Latent vs explicit knowledge representation: how ChatGPT answers questions about low-frequency entities. In: Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024). ELRA and ICCL, Torino, Italia, pp 10172–10185.

Graff D (2002) The AQUAINT corpus of English news text. Linguistic Data Consortium. https://catalog.ldc.upenn.edu/LDC2002T31

Gu Y, Dong L, Wei F, Huang, M (2024) MiniLLM: knowledge distillation of large language models. In: The twelfth international conference on learning representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net. https://openreview.net/forum?id=5h0qf7IBZZ

Guellil I, Garcia-Dominguez A, Lewis PR, Hussain S, Smith G (2024) Entity linking for English and other languages: a survey. Knowl Inf Syst 66(7):3773–3824. https://doi.org/10.1007/s10115-023-02059-2

Hamdi A, Linhares Pontes E, Boros E, Nguyen TTH, Hackl G, Moreno JG, Doucet A (2021) A multilingual dataset for named entity recognition, entity linking and stance detection in historical newspapers. In: Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. SIGIR '21. Association for Computing Machinery, New York, NY, USA, pp 2328–2334. https://doi.org/10.1145/3404835.3463255

Hayes AF, Krippendorff K (2007) Answering the call for a standard reliability measure for coding data. Commun Methods Meas 1(1):77–89. https://doi.org/10.1080/19312450709336664

Hoffart J, Yosef MA, Bordino I, Fürstenau H, Pinkal M, Spaniol M, Taneva B, Thater S, Weikum G (2011) Robust disambiguation of named entities in text. In: Proceedings of the 2011 conference on empirical methods in natural language processing. Association for Computational Linguistics, Edinburgh, Scotland, UK, pp 782–792. https://aclanthology.org/D11-1072

Hoffart J, Seufert S, Nguyen DB, Theobald M, Weikum G (2012) KORE: keyphrase overlap relatedness for entity disambiguation. In: Proceedings of the 21st ACM international conference on information and knowledge management. CIKM '12. Association for Computing Machinery, New York, NY, USA, pp 545–554. https://doi.org/10.1145/2396761.2396832

Hummel RA, Zucker SW (1983) On the foundations of relaxation labeling processes. IEEE Trans Pattern Anal Mach Intell PAMI-5(3):267–287. https://doi.org/10.1109/TPAMI.1983.4767390

Ilievski F, Vossen P, Schlobach S (2018) Systematic study of long tail phenomena in entity linking. In: Bender EM, Derczynski L, Isabelle P (eds) Proceedings of the 27th international conference on computational linguistics. Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp 664–674. https://aclanthology.org/C18-1056

Jacovi A, Caciularu A, Goldman O, Goldberg Y (2023) Stop uploading test data in plain text: practical strategies for mitigating data contamination by evaluation benchmarks. In: Bouamor H, Pino J, Bali K (eds) Proceedings of the 2023 conference on empirical methods in natural language processing. Association for Computational Linguistics, Singapore, pp 5075–5084. https://doi.org/10.18653/v1/2023.emnlp-main.308, https://aclanthology.org/2023.emnlp-main.308

Ji H, Grishman R, Dang HT, Griffit K, Ellis J (2010) Overview of the tac 2010 knowledge base population track. In: Proceedings of the 2010 text analysis conference

Jurafsky D, Martin JH (2023) Speech and language processing, (2023). Draft of January 7. https://web.stanford.edu/$~$jurafsky/slp3/. Accessed 13 Aug 2024

Kandpal N, Deng H, Roberts A, Wallace E, Raffel C (2023) Large language models struggle to learn long-tail knowledge. In: Proceedings of the 40th international conference on machine learning. ICML'23. JMLR.org, Honolulu, Hawaii, USA

Kolitsas N, Ganea O-E, Hofmann T (2018) End-to-end neural entity linking. In: Korhonen A, Titov I (eds) Proceedings of the 22nd conference on computational natural language learning. Association for Computational Linguistics, Brussels, Belgium, pp 519–529. https://doi.org/10.18653/v1/K18-1050

Lacerra C, Pasini T, Tripodi R, Navigli R (2021a) ALaSca: an automated approach for large-scale lexical substitution. In: Proceedings of the thirtieth international joint conference on artificial intelligence (IJCAI-21). International Joint Conferences on Artificial Intelligence Organization, Montreal, Canada, pp 3836–3842. https://doi.org/10.24963/ijcai.2021/528

Lacerra C, Tripodi R, Navigli R (2021b) GeneSis: a generative approach to substitutes in context. In: Proceedings of the 2021 conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp 10810–10823. https://doi.org/10.18653/v1/2021.emnlp-main.844

Li C, Flanigan J (2024) Task contamination: language models may not be few-shot anymore. Proc AAAI Conf Artif Intell 38(16):18471–18480. https://doi.org/10.1609/aaai.v38i16.29808

Mallen A, Asai A, Zhong V, Das R, Khashabi D, Hajishirzi H (2023) When not to trust language models: investigating effectiveness of parametric and non-parametric memories. In: Rogers A, Boyd-Graber J, Okazaki N (eds) Proceedings of the 61st annual meeting of the Association for Computational Linguistics (volume 1: long papers). Association for Computational Linguistics, Toronto, Canada, pp 9802–9822. https://doi.org/10.18653/v1/2023.acl-long.546

Menzel S, Schnaitter H, Zinck J, Petras V, Neudecker C, Labusch K, Leitner E, Rehm G (2021) In: Franke-Maier M, Kasprzik A, Ledl A, Schürmann H (eds) Named Entity Linking mit Wikidata und GND - Das Potenzial handkuratierter und strukturierter Datenquellen für die semantische Anreicherung von Volltexten. De Gruyter Saur, Berlin, Boston, pp 229–258. https://doi.org/10.1515/9783110691597-012

Miller GA (1995) WordNet: a lexical database for English. Commun ACM 38(11):39–41. https://doi.org/10.1145/219717.219748

Miller DA, Zucker SW (1991) Copositive-plus Lemke algorithm solves polymatrix games. Oper Res Lett 10(5):285–290. https://doi.org/10.1016/0167-6377(91)90015-H

Mitchell A, Strassel S, Huang S, Zakhary R (2005) ACE 2004 multilingual training corpus LDC2005T09. Linguistic Data Consortium, Philadelphia

Möller C, Lehmann J, Usbeck R (2022) Survey on English entity linking on Wikidata: datasets and approaches. Semant Web 13(6):925–966. https://doi.org/10.3233/SW-212865

Nash J (1951) Non-cooperative games. Ann Math 54:286–295

Orlando R, Huguet Cabot P-L, Barba E, Navigli R (2024) ReLiK: Retrieve and LinK, fast and accurate entity linking and relation extraction on an academic budget. In: Ku L-W, Martins A, Srikumar V (eds) Findings of the Association for Computational Linguistics ACL 2024. Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, pp 14114–14132. https://aclanthology.org/2024.findings-acl.839

Pelillo M (1997) The dynamics of nonlinear relaxation labeling processes. J Math Imaging Vis 7(4):309–323. https://doi.org/10.1023/A:1008255111261

Petroni F, Piktus A, Fan A, Lewis P, Yazdani M, De Cao N, Thorne J, Jernite, Y, Karpukhin V, Maillard J, Plachouras V, Rocktäschel T, Riedel S (2021) KILT: a benchmark for knowledge intensive language tasks. In: Proceedings of the 2021 conference of the North American chapter of the Association for Computational Linguistics: human language technologies. Association for Computational Linguistics, Online, pp 2523–2544. https://doi.org/10.18653/v1/2021.naacl-main.200

Plekhanov M, Kassner N, Popat K, Martin L, Merello S, Kozlovskii B, Dreyer FA, Cancedda N (2023) Multilingual end to end entity linking. arXiv Preprint http://arxiv.org/abs/2306.08896

Pontes EL, Cabrera-Diego LA, Moreno JG, Boros E, Hamdi A, Doucet A, Sidère N, Coustaty M (2022) MELHISSA: a multilingual entity linking architecture for historical press articles. Int J Digit Libr 23:133–160

Provatorova V, Bhargav S, Vakulenko S, Kanoulas E (2021) Robustness evaluation of entity disambiguation using prior probes: the case of entity overshadowing. In: Moens M-F, Huang X, Specia L, Yih SW-t (eds) Proceedings of the 2021 conference on empirical methods in natural language processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp 10501–10510. https://doi.org/10.18653/v1/2021.emnlp-main.820

Ramirez-Orta JA, Xamena E, Maguitman A, Milios E, Soto AJ (2022) Post-OCR document correction with large ensembles of character sequence-to-sequence models. In: Proceedings of the AAAI conference on artificial intelligence. AAAI Press, pp 11192–11199. https://doi.org/10.1609/aaai.v36i10.21369

Rota Bulò S, Pelillo M, Bomze IM (2011) Graph-based quadratic optimization: a fast evolutionary approach. Comput Vis Image Underst 115(7):984–995. https://doi.org/10.1016/j.cviu.2010.12.004. (**Special issue on Graph-Based Representations in Computer Vision**)

Sainz O, Campos J, García-Ferrero I, Etxaniz J, Lacalle OL, Agirre E (2023) NLP evaluation in trouble: on the need to measure LLM data contamination for each benchmark. In: Bouamor H, Pino J, Bali K (eds) Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, Singapore, pp 10776–10787. https://doi.org/10.18653/v1/2023.findings-emnlp.722, https://aclanthology.org/2023.findings-emnlp.722

Santini C, Tan MA, Bruns O, Tietz T, Posthumus E, Sack H (2022) Knowledge extraction for art history: the case of Vasari's the lives of the artists. In: Paschke A, Rehm G, Neudecker C, Pintscher L (eds) Proceedings of the third conference on digital curation technologies (Qurator 2022). CEUR Workshop Proceedings, vol 3234. CEUR-WS.org, Berlin, Germany. https://ceur-ws.org/Vol-3234/paper7.pdf

Scarlini B, Pasini T, Navigli R (2020) With more contexts comes better performance: contextualized sense embeddings for all-round word sense disambiguation. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics, Online, pp 3528–3539. https://doi.org/10.18653/v1/2020.emnlp-main.285

Sevgili Ö, Shelmanov A, Arkhipov M, Panchenko A, Biemann C (2022) Neural entity linking: a survey of models based on deep learning. Semant Web 13(3):527–570. https://doi.org/10.3233/SW-222986

Sun K, Xu YE, Zha H, Liu Y, Dong XL (2024) Head-to-tail: how knowledgeable are large language models (LLMs)? A.K.A. will LLMs replace knowledge graphs? arXiv

Taylor PD, Jonker LB (1978) Evolutionary stable strategies and game dynamics. Math Biosci 40(1):145–156. https://doi.org/10.1016/0025-5564(78)90077-9

Tedeschi S, Conia S, Cecconi F, Navigli R (2021) Named entity recognition for entity linking: what works and what's next. In: Moens M-F, Huang X, Specia L, Yih SW-t (eds) Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics, Punta Cana, Dominican Republic, pp 2584–2596. https://doi.org/10.18653/v1/2021.findings-emnlp.220

Tripodi R, Navigli R (2019) Game theory meets embeddings: a unified framework for word sense disambiguation. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, pp 88–99. https://doi.org/10.18653/v1/D19-1009

Tripodi R, Blloshmi R, Levis Sullam S (2022) Evaluating multilingual sentence representation models in a real case scenario. In: Proceedings of the thirteenth language resources and evaluation conference. European Language Resources Association, Marseille, France, pp 2928–2939. https://aclanthology.org/2022.lrec-1.314

van Hulst JM, Hasibi F, Dercksen K, Balog K, Vries AP (2020) REL: an entity linker standing on the shoulders of giants. In: Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval. SIGIR '20. Association for Computing Machinery, New York, NY, USA, pp 2197–2200. https://doi.org/10.1145/3397271.3401416

Wang J, Dong Y (2020) Measurement of text similarity: a survey. Information 11(9):421. https://doi.org/10.3390/INFO11090421

Weibull JW (1997) Evolutionary game theory. MIT Press, Cambridge

Wu L, Petroni F, Josifoski M, Riedel S, Zettlemoyer L (2020) Scalable zero-shot entity linking with dense entity retrieval. In: Webber B, Cohn T, He Y, Liu Y (eds) Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics, Online, pp 6397–6407. https://doi.org/10.18653/v1/2020.emnlp-main.519

Zaporojets K, Kaffee L-A, Deleu J, Demeester T, Develder C, Augenstein I (2022) TempEL: linking dynamically evolving and newly emerging entities. In: Thirty-sixth conference on neural information processing systems datasets and benchmarks track. https://openreview.net/forum?id=vrnqr3PG4yB

Zhang W, Hua W, Stratos K (2022) EntQA: entity linking as question answering. In: International conference on learning representations. https://openreview.net/forum?id=US2rTP5nm_

Zhu F, Yu J, Jin H, Hou L, Li J, Sui Z (2023) Learn to not link: exploring NIL prediction in entity linking. In: Rogers A, Boyd-Graber J, Okazaki N (eds) Findings of the Association for Computational Linguistics: ACL 2023. Association for Computational Linguistics, Toronto, Canada, pp 10846–10860. https://doi.org/10.18653/v1/2023.findings-acl.690

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.