International Journal on Digital Libraries (2022) 23:133–160
https://doi.org/10.1007/s00799-021-00319-6

MELHISSA: a multilingual entity linking architecture for historical press articles

Elvys Linhares Pontes[1] · Luis Adrián Cabrera-Diego[2] · Jose G. Moreno[3] · Emanuela Boros[2] · Ahmed Hamdi[2] · Antoine Doucet[2] · Nicolas Sidere[2] · Mickaël Coustaty[2]

Received: 1 May 2021 / Revised: 27 October 2021 / Accepted: 30 October 2021 / Published online: 29 November 2021

Abstract

Digital libraries have a key role in cultural heritage as they provide access to our culture and history by indexing books and historical documents (newspapers and letters). Digital libraries use natural language processing (NLP) tools to process these documents and enrich them with meta-information, such as named entities. Despite recent advances in these NLP models, most of them are built for specific languages and contemporary documents that are not optimized for handling historical material that may for instance contain language variations and optical character recognition (OCR) errors. In this work, we focused on the entity linking (EL) task that is fundamental to the indexation of documents in digital libraries. We developed a Multilingual Entity Linking architecture for HIstorical preSS Articles that is composed of multilingual analysis, OCR correction, and filter analysis to alleviate the impact of historical documents in the EL task. The source code is publicly available. Experimentation has been done over two historical document corpora covering five European languages (English, Finnish, French, German, and Swedish). Results have shown that our system improved the global performance for all languages and datasets by achieving an F-score@1 of up to 0.681 and an F-score@5 of up to 0.787.

Keywords Entity linking · Historical data · Digital libraries · Deep learning · Heuristics

1 Introduction

Historical documents are an essential resource in the understanding of our cultural heritage. The development of recent technologies, such as optical character recognition (OCR) systems, eases the digitization of physical documents and the extraction of textual content. Digitization provides two major advantages, in particular for digital humanities (DH) scholars: the exponential increase of target audiences, and the preservation of original documents from any damage when accessing them [1–4]. The recent interest in massive digitization raises multiple challenges to content providers including indexing, categorization, searching, to mention a few. Although these challenges also exist when dealing with contemporary text documents, digitized documents make them harder because of inherent problems associated with the source quality (natural degradation of the documents) and to the digitization process itself (e.g. digitization noise, image quality, and OCR bias) [5–11].

Digitizing historical documents does not only increase the availability of these resources but also allows digital humanities researchers to search, structure, and organize

✉ Antoine Doucet
antoine.doucet@univ-lr.fr

Elvys Linhares Pontes
elvyslpontes@gmail.com

Luis Adrián Cabrera-Diego
luis.cabrera_diego@univ-lr.fr

Jose G. Moreno
jose.moreno@irit.fr

Emanuela Boros
emanuela.boros@univ-lr.fr

Ahmed Hamdi
ahmed.hamdi@univ-lr.fr

Nicolas Sidere
nicolas.sidere@univ-lr.fr

Mickaël Coustaty
mickael.coustaty@univ-lr.fr

[1] Trading Central Labs, Sophia Antipolis, France

[2] L3i, La Rochelle Université, La Rochelle, France

[3] IRIT, University of Toulouse, Toulouse, France

information located within the documents [1,2,12]. For instance, researchers might use digitized documents to identify tangible keywords (i.e. people, places, events) but also more abstract, varied, and subtler concepts, such as themes and topics. Furthermore, digitized historical documents have allowed the use of natural language processing (NLP) tools, such as named entity recognition (NER) [9–11] and entity linking (EL) [5,7] for enriching automatically the documents. This has attracted the attention of numerous digital humanities researchers since it allows quantitative analysis, e.g. towards finding patterns in historical documents on cultural changes, variations in gender bias across historical periods, emerging technological trends, or transitions to new political ideas [3,4].

Despite the interest of digital humanities researchers in NLP and information retrieval (IR) tools, the creation of these for processing contemporary and historical documents has been disproportionate. For contemporary documents, in the last decade, the number of tools has increased until the point where they have been generally adopted. However, this has not been the case for historical documents, due to certain characteristics, which make their processing particularly difficult. A few exceptions exist [9,13,14], but in far smaller numbers than for contemporary documents. Among the challenges, such tools need to be able to deal with errors produced by OCR systems, to manage some specific vocabulary, and also to handle spelling variations with respect to modern standards. To ease the impact of OCR errors, one solution is to apply post-OCR correction [15,16], but while beneficial, this process remains imperfect.

To illustrate and extend some of the aforementioned problems, we present in Fig. 1 a collection of images representing historical newspapers or portions of them. As we can observe in Fig. 1a–c, newspapers can have different templates but also face an unbalanced level of degradation. In the case of Fig. 1c, d we can observe a stamp that covers parts of the original text and makes portions of it illegible. Figure 1e provides an example of a text containing a word that is nowadays spelled differently, which makes it difficult to match with contemporary knowledge bases. In Fig. 1f–g, we present two fonts that can be difficult to process by an OCR system due to the geometry of certain characters, such as 𝔖 (S), 𝔓 (P), and ſ (Long S). For instance, in Fig. 1g, the word "Conſtitution" was recognized as "Conftitution" by an OCR system[1]. Finally, in Fig. 1h, we present a document where we can notice a mix between French and English within a single document.

Apart from digitizing and recognizing the text, the processing of historical documents consists in extracting metadata from them. This metadata is used to index the key information inside documents to ease the navigation and retrieval process. Among all the possible key information available, named entities are of major significance as they allow structuring the document content [17], and correspond to key elements queried for in search engines [18]. These entities can represent aspects such as people, places, organizations, and events. Nonetheless, historical documents may contain duplicated and ambiguous information about named entities due to the heterogeneity and the mix of temporal references [19,20]. A disambiguation process is thus essential to distinguish named entities to be further utilized by search systems in digital libraries. For instance, "Bonaparte" can refer to several entities: the general "Napoleon Bonaparte"[2] or his son, "Napoleon François Joseph Charles Bonaparte"[3], but also a German band[4], to name a few.

Entity linking (EL) aims to recognize, disambiguate, and relate named entities to specific entries in a knowledge base. EL is a challenging task due to the fact that named entities may have multiple surface forms, for instance, in the case of a person an entity can be represented with their full or partial name, alias, honorifics, or alternate spellings [21]. Compared to contemporary data, few works in the state of the art have studied the EL task on historical documents [19,20,22–26] and OCR-processed documents [5].

In our previous work [27], we proposed a combination of a multilingual end-to-end entity linking method with several techniques to minimize the impact of issues frequently found in historical data. Our EL approach made use of entity embeddings, built from Wikipedia in multiple languages, along with a neural attention mechanism that analyzes context words and candidate entity embeddings to disambiguate mentions in historical documents. To adjust to historical documents, we developed several modules to handle multilingualism and errors stemming from the output of OCR systems.

In this paper, we present MELHISSA, a Multilingual Entity Linking architecture for HIstorical preSS Articles, which extends our previous work on EL [27]. Specifically, we present an EL analysis on two recent historical datasets: CLEF HIPE 2020 [13] and NewsEye [28] that are composed of documents in English, Finnish, French, German, and Swedish. This deep analysis enabled us to improve our approach and achieve better results for both datasets and all languages.

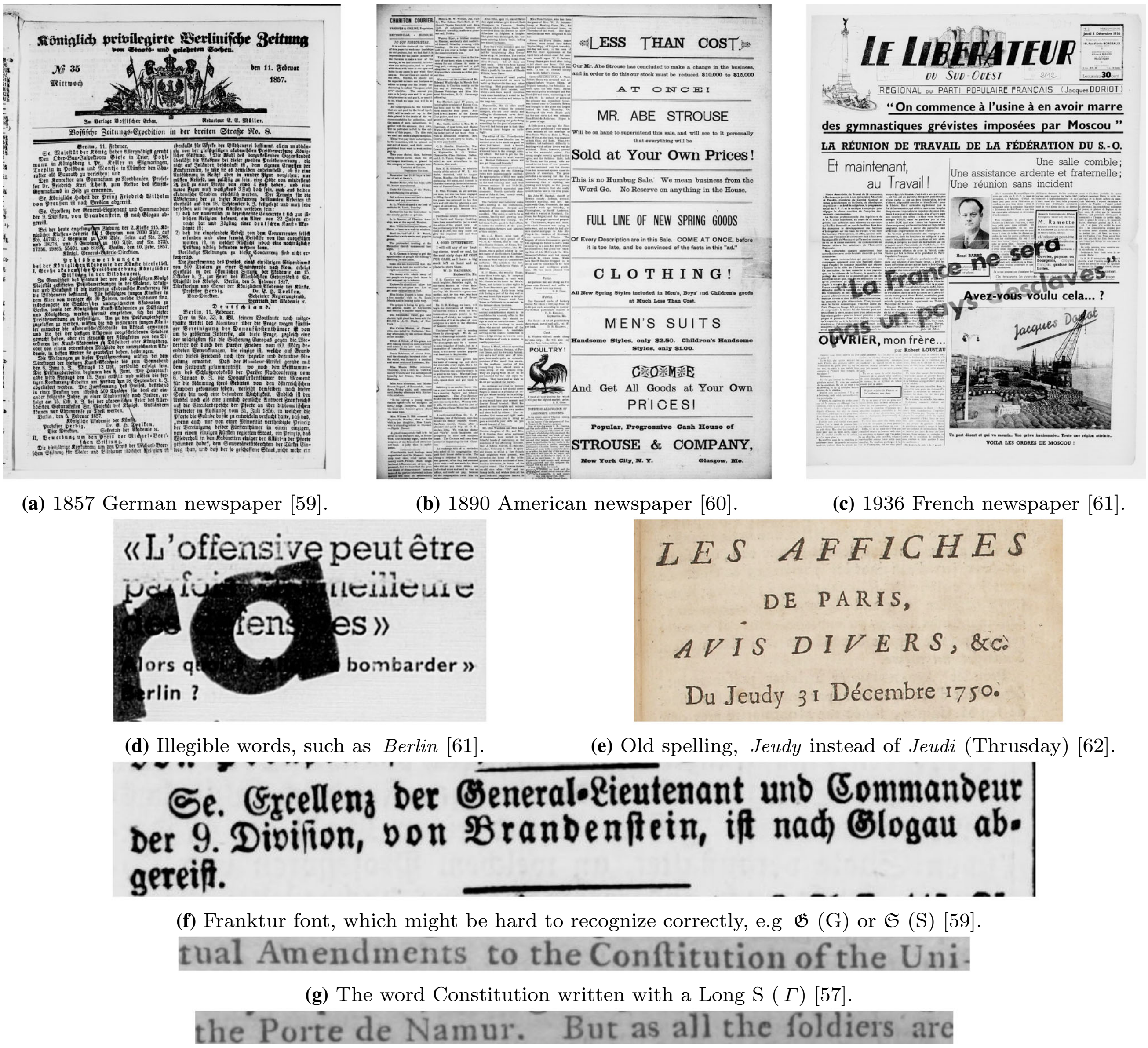
This paper is organized as follows. We present an overview of EL approaches and a survey on historical data for the EL task in Sect. 2, before describing our multilingual approach in Sect. 3. Next, the CLEF HIPE 2020 and NewsEye datasets are described in Sect. 4. Then, the experimental setup is introduced in Sect. 5, while the results are presented in Sect. 6. We discuss the results in Sect. 7. Finally, we provide conclusions and final comments in Sect. 8.

[1] HIPE-data-v1.3-test-masked-bundle5-en.tsv#L56-L61

[2] https://www.wikidata.org/wiki/Q517.

[3] https://www.wikidata.org/wiki/Q7723.

[4] https://www.wikidata.org/wiki/Q892094.


Königlich privilegirte Berliniſche Zeitung
LESS THAN COST
MR. ABE STROUSE
Sold at Your Own Prices!
CLOTHING!
STROUSE & COMPANY.
LE LIBÉRATEUR
DU SUD-OUEST
"La France ne sera pas un pays d'esclaves"

(a) 1857 German newspaper [59]. **(b)** 1890 American newspaper [60]. **(c)** 1936 French newspaper [61].

«L'offensive peut être
lors q
bombarder »
rlin ?

(d) Illegible words, such as *Berlin* [61].

LES AFFICHES
DE PARIS,
AVIS DIVERS, &c.
Du Jeudy 31 Décembre 1750.

(e) Old spelling, *Jeudy* instead of *Jeudi* (Thrusday) [62].

Se. Excellenz der General-Lieutenant und Commandeur der 9. Diviſion, von Brandenſtein, iſt nach Glogau abgereiſt.

(f) Franktur font, which might be hard to recognize correctly, e.g 𝔊 (G) or 𝔖 (S) [59].

tual Amendments to the Conſtitution of the Uni-

(g) The word Constitution written with a Long S (ſ) [57].

the Porte de Namur. But as all the ſoldiers are

(h) Use of a name location in French, *Porte de Namur* (Namur Gate), within an English document [58].

Fig. 1 Examples of historical newspaper documents [57–62]

2 Entity linking for historical data

Entity linking (EL) is an information extraction (IE) task that semantically enriches documents by identifying pieces of text that refer to entities, generally depicted as mention detection, and by matching each piece to an entry in a knowledge base (KB). Frequently, the detection of mentions is delegated to an external named entity recognition (NER) system. In the state of the art of EL, the systems are either disambiguation systems [29,30], i.e. tools that perform only the matching of entities and consider the first task as an input, or end-to-end systems [22,25,26,31–33], i.e. tools that jointly perform both tasks, detecting and disambiguating the entities at the same time.

In the last year, new methods have been proposed for disambiguating entities and to solve specific issues, such as domain overfitting and context neglection. For instance, Onoe and Durrett [30] proposed a disambiguation system to overcome the risk of EL methods of overfitting to the domain (the genre of text or the particular distribution of entities), and, in consequence, to generalize effectively. The model does not rely on labelled entity linking data with a specific entity distribution. The authors derive a large inventory of types from Wikipedia categories and use hyperlinked mentions in Wikipedia to distantly label data and train an entity typing model. With this domain-independent setting, their approach achieves strong results on the CoNLL dataset [34].

While most disambiguation systems employ entity representations embeddings bootstrapped from word embeddings to assess topic-level context compatibility, they also tend to neglect the context of the mention. A recent method, [35], injects latent entity type information into the entity embeddings based on the widely utilized pre-trained bidirectional encoder representations from Transformers (BERT) [36]. Then, it integrates a BERT-based entity similarity score into the local context model of a state-of-the-art model to better capture latent entity type information. This method significantly outperformed the state-of-the-art entity linking models on the standard benchmark (AIDA-CoNLL [37]).

The other main type of systems, end-to-end EL systems, were initially defined for modern documents [32]. However, as time passed, researchers have considered end-to-end EL systems also for historical documents [38]. Furthermore, the first end-to-end EL systems were focused on monolingual corpora and have gradually moved to cross-lingual and multilingual contexts. For example, a recent configuration, cross-lingual named entity linking (XEL), consists of analyzing documents and named entities in a language different from the one used in the knowledge base (KB). Several recent works proposed different XEL approaches: zero-shot transfer learning method by using a pivot language [39], a hybrid approach using language-agnostic features that combine existing lookup-based and neural candidate generation methods [40], and the use of multilingual word embeddings to disambiguate mentions across languages [7].

Another work [41] proposed a new approach to assess the problems faced by their previous entity candidate generation methods [40] for low-resource XEL. They reduce the disconnection between entity mentions and KB entries by introducing mention-entity pairs into the training process to provide supervision. Also, their approach improves the robustness of the model to low-resource scenarios by adjusting their previous neural-based model.

Further, an end-to-end BERT-based system [33] was advocated for EL by casting a token classification over the entire entity vocabulary (an entity vocabulary, in this case, would be of a considerably large amount, e.g. 700k). The authors showed on an entity linking benchmark that this improved the entity representations over plain BERT and that it outperformed EL architectures that optimized the tasks separately.

In Digital Humanities, EL systems dedicated to historical documents have also been explored [22,25,26,42]. For instance, van Hooland et al. [22] evaluated three third-party entity extraction services through a comprehensive case study, based on the descriptive fields of the Smithsonian Cooper-Hewitt National Design Museum in New York. Ruiz and Poibeau [26] utilized the DBpedia Spotlight tool[5] to disambiguate named entities on Bentham manuscripts[6]. Moreover, Munnelly and Lawless [42] investigated the accuracy and overall suitability of EL systems in Seventeenth-century depositions obtained during the 1641 Irish Rebellion[7].

Most of the developed EL systems in Digital Humanities are monolingual. Several disambiguation systems have been studied by focusing on specific types of entities in historical documents, e.g. person and place names. Smith and Crane [19] investigated the identification and disambiguation of place names in the Perseus digital library. They concentrated on representing historical data in the humanities from Ancient Greece to Nineteenth century America. In order to overcome the heterogeneous data and the mix of temporal references (e.g. places that changed name over time), they proposed a method based on honorifics, generic geographic labels, and linguistic environments to recognize entities, while they made use of gazetteers, biographical information, and general linguistic knowledge to disambiguate these entities. Other works [23,24] focused on author names in French literary criticism texts and scientific essays from the 19th and early 20th centuries. They proposed a graph-based method that leverages knowledge from different linked data sources to generate the list of candidates for each author mention. It then crawls data from other linked datasets using equivalence links and fuses graphs of homologous individuals into a non-redundant graph in order to select the best candidate.

Dedicated end-to-end EL systems for historical documents have also been developed [22,25,26]. Some concentrated on developing features and rules for improving EL in a specific domain [20], while others focused on efficiently utilizing entity types [19,23,24]. Furthermore, some researchers investigated the effect of the issues frequently found in historical documents on the task of EL [5,20]. Most of the proposed systems were also monolingual. The work of Mosallam *et al.* [38] proposed a monolingual unsupervised method to recognize person names, locations, and organizations in digitized French journals of the National Library of France (Bibliothèque nationale de France) from the Nineteenth century. Then, they used a French entity knowledge base along with a statistical contextual disambiguation approach. Interestingly, their method outperformed supervised approaches when trained on small amounts of annotated data. Huet et al. [43] also analyzed the French journal Le Monde archive, a collection of documents from 1944 until 1986 discussing different subjects (e.g. post-war period, end of colonialism, politics, sports, culture). The authors calculated a conditional distribution of the co-occurrence of mentions with their cor-

[5] https://www.dbpedia-spotlight.org/.

[6] https://www.ucl.ac.uk/library/digital-collections/collections/bentham.

[7] https://1641.tcd.ie/.

responding entities (Wikipedia article). Then, they linked these Wikipedia articles to YAGO [44] to recognize and disambiguate entities in the archive of Le Monde.

Heino et al. [20] investigated EL in a particular domain, the Second World War in Finland, using the reference datasets of WarSampo[8]. They proposed a ruled-based approach to disambiguate military units, places, and people in these datasets. Moreover, they investigated problems regarding the analysis and disambiguation of these entities in this kind of data, while they proposed specific rules to overcome these issues.

Regarding the lack of resources in the context of Digital Humanities, a recent study explored the low resource settings with costly annotated data and domain-specific KBs [45]. The approach proposes a domain-agnostic feedback-based annotation approach based on suggestions from the annotators of potential concepts and adaptive candidate ranking. The method improves the annotation process by 35% compared to annotating without interactive support.

EL in historical datasets relies on information such as names or locations that are both non-unique and prone to enumeration and transcription errors. These errors make it impossible to find the correct match with certainty. A recent paper [46] brings forward a fully automated probabilistic method for linking historical datasets that enable researchers to create samples at the frontier of minimizing false positives and false negatives, by utilizing the expectation-maximization (EM) algorithm. The authors study the method to link historical population censuses in the US and Norway and use these samples to estimate measures of intergenerational occupational mobility.

The impact of OCR errors on EL systems, to our knowledge, has rarely been analyzed or alleviated in previous research. Thus, the ability of EL to handle noisy inputs continuous to be an open question. Nevertheless, Linhares Pontes et al. [5], reported that EL systems for contemporary documents can see their performance decreasing around 20% when OCR errors, at the character and word levels, reach rates of 5% and 15%, respectively.

Differently from previous works, we propose a multilingual end-to-end approach to link entities mentioned in historical documents to a KB containing several techniques to reduce the impact of the issues generated by the historical data issues, e.g. multilingualism, grammatical errors generated by OCR engines, linguistic historical word variations. The next section details our approach.

3 Multilingual end-to-end entity linking

As aforementioned, historical documents present particular characteristics that make EL particularly challenging. In the following subsections, we describe the methods and techniques we developed for creating *MELHISSA*, our EL system that addresses these challenges.

3.1 Building resources

The main component of an EL system is its knowledge base (KB) which allows the storage of the full list of entities used as reference. Modern KBs are rich enough to deal with additional tasks such as extraction of supplementary contexts or surface names, disambiguation of cases, or linking of entities with a particular website entry. A well-known set of publicly available KBs are Wikipedia[9], Wikidata[10], and DBpedia [47]. Here, we briefly describe these KBs.

Wikipedia is a multilingual encyclopaedia that includes more than 300 languages, but only near to 70 languages have more than 100,000 articles. It is a widely used KB as a source of information for EL systems but also for building datasets. Multiple research studies, e.g. [29,31,48], make use of the English Wikipedia to train their models and disambiguate entity mentions. However, it has also been used to study the matching of mentions to Wikipedia articles based exclusively on their cultural heritage as well as for the disambiguation of mentions found in historical documents [49].

Wikidata is a KB created by the Wikimedia Foundation[11]. Its main purpose is to store user-generated data from the various projects supported by Wikimedia. Wikidata is widely used as a standard reference for entities, in the context of digital humanities, Wikidata has been used to annotate CLEF HIPE 2020 and NewsEye, two EL datasets of historical documents.

DBpedia is a KB that categorizes data from different Wikimedia projects, like Wikipedia and Wikidata. Furthermore, it associates this information to other KBs such as YAGO [44] and GeoNames[12]. It has been used in different projects related to EL [22,25,50,51]. For instance, De Wilde [51] used it for linking locations in a historical newspaper corpus and Munnelly and Lawless [25] utilized DBpedia for annotating historical legal documents.

In contrast with the aforementioned research [22,25,26], we built our own domain-independent KB mainly based on Wikipedia. In order to cover a large number of languages and long-tail entities, we made use of the Wikipedia versions of our target languages, e.g. French, German, Finnish, and Swedish, as well as the English Wikipedia. Our idea behind this strategy is that despite the richness and coverage of the English Wikipedia, in some cases other versions of Wikipedia might contain information that is only found in a

[8] https://seco.cs.aalto.fi/projects/sotasampo/en/.

[9] https://www.wikipedia.org.

[10] https://www.wikidata.org.

[11] https://www.wikimedia.org.

[12] http://www.geonames.org.

specific language. This situation is less frequent for popular entities but common when dealing with long-tail entities. For instance, *Maurice Maréchal*, a journalist and founder of the French newspaper *Le Canard enchaîné*, has entries only in the French and Esperanto Wikipedias[13].

3.2 Probabilistic table entity map

In order to provide relevant candidates for mentions, for each explored language we downloaded the last version of their Wikipedia dump (as of March 2021), and analyzed its pages. We collected all hyperlinks presented in these pages to map the Wikipedia pages to their surface variation names represented in the hyperlinks. In Table 1, we present the most representative statistics regarding Wikipedia dumps (1a) and their processed entities and surface names (1b). It can be observed in Table 1a and their processed entities are that the number of entities used in our KBs is greater than the number of articles found in each Wikipedia dump. The reason for this is that our KBs also include pages that refer to redirections and disambiguation pages.

Figure 2 shows examples extracted from the English Wikipedia. While most mentions (in blue colour and underlined) have the same surface representation as their links to the Wikipedia pages (e.g. “Association football”, “1904 Summer Olympics”, “St. Louis”, “Canada”, and “Ontario”), the mention “United States” represents the Olympic and Paralympic committee of United States which has a shorter surface representation of its mention. The mention United States can also represent the country in Fig. 2b. Finally, Fig. 2c shows an example where the mention “United States of America” is longer than its entity (“United States”). It should be indicated that surface name “United States” can be linked to 637 entities, while the mention “United States of America” can be related to 28 entities.

From these maps, we calculate the probability of an entity (Wikipedia page) e to be related to a mention (surface representation) m:

$$p(e|m) = \frac{|m \mapsto e|}{|m|} \tag{1}$$

where $|m \mapsto e|$ is the number of times that mention m refers to e within Wikipedia and $|m|$ is the total number of occurrences of the mention m in the Wikipedia dump. From this probabilistic table, it is possible to find which are the top entities that a mention span refers to. For instance, the mention “United States” has a probability of 95.9% to be related to the entity “United States” and 1×10^{-6} to the entity “United States Olympic & Paralympic Committee”.

Table 1 Statistics regarding the data used in this work

(a) Wikipedia dump. The number of entities correspond to the number of pages used in our KBs

Lang	Pages (M)	Articles (M)	Entities (M)
de	7.1	2.5	2.8
en	53	6.2	7.9
fi	1.3	0.5	0.5
fr	11	2.3	2.6
sv	7.1	3.2	3.7

(b) Number of entities associated per surface name (mention). Between brackets, we provide statistics for the most frequent 1,000 entities

Lang	Min	1st Quantile	Median	Mean	3rd Quantile	Max
de	1 (1)	1.00 (6.00)	1 (18)	1.19 (46.31)	1.00 (44.25)	982 (982)
en	1 (1)	1.00 (13.75)	1 (42)	1.23 (100.12)	1.00 (95.00)	1734 (1734)
fi	1 (1)	1.00 (1.00)	1 (1)	1.10 (7.39)	1.00 (5.00)	140 (110)
fr	1 (1)	1.00 (9.00)	1 (20)	1.235 (55.68)	1.00 (50.00)	1038 (1038)
sv	1 (1)	1.00 (1.00)	1 (3)	1.124 (11.26)	1.00 (11.00)	237 (162)

[13] https://fr.wikipedia.org/wiki/Maurice_Mar%C3%A9chal_(journaliste).

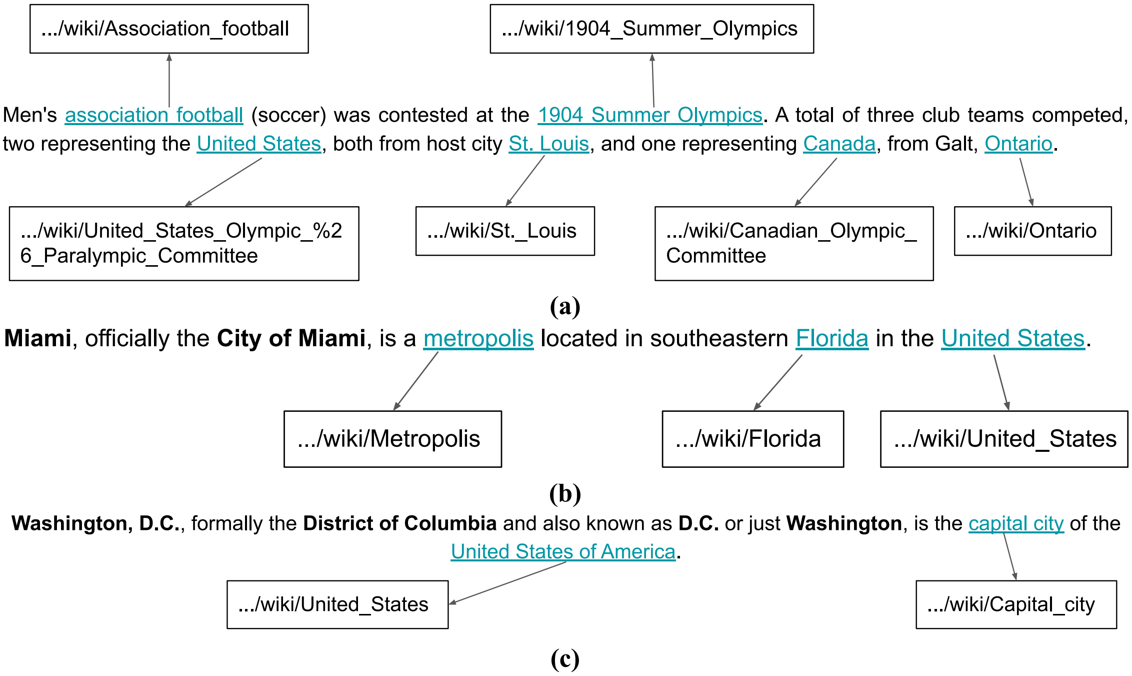

.../wiki/Association_football
.../wiki/1904_Summer_Olympics
Men's association football (soccer) was contested at the 1904 Summer Olympics. A total of three club teams competed, two representing the United States, both from host city St. Louis, and one representing Canada, from Galt, Ontario.
.../wiki/United_States_Olympic_%26_Paralympic_Committee
.../wiki/St._Louis
.../wiki/Canadian_Olympic_Committee
.../wiki/Ontario
(a)
Miami, officially the City of Miami, is a metropolis located in southeastern Florida in the United States.
.../wiki/Metropolis
.../wiki/Florida
.../wiki/United_States
(b)
Washington, D.C., formally the District of Columbia and also known as D.C. or just Washington, is the capital city of the United States of America.
.../wiki/United_States
.../wiki/Capital_city
(c)

Fig. 2 Examples of mentions and their links to Wikipedia pages. Sentences extracted from the Wikipedia pages: https://en.wikipedia.org/wiki/Football_at_the_1904_Summer_Olympics, https://en.wikipedia.org/wiki/Miami, and https://en.wikipedia.org/wiki/Washington,_D.C.

3.3 Entity embeddings

We create entity embeddings for each language, in the same manner as in [29], by generating two conditional probability distributions:

- The *positive probability distribution* is an approximation based on the word-entity co-occurrence counts, i.e. which words appear in the context of an entity. These counts were obtained from the entity Wikipedia pages, and from the surrounding context of the entity in the corpus, by utilizing a fixed-length window.
- The *negative probability distribution* is calculated by the random sampling of context windows that were unrelated to a specific entity.

These probability distributions were utilized with the purpose of changing the alignment of the word embeddings with respect to an entity embedding. While the *positive probability distribution* should approach the embeddings of the co-occurring words with the entity embedding, *the negative probability distribution* should distance the word embeddings that affiliated or related to an entity.

In order to prevent bias and low generalization, we create these word embeddings without relying or depending on the dataset. In the case where an entity does not have entity embeddings, the EL system will assign it to NIL, meaning that it could not find a reference that it considers likely enough to be adequate. This may actually be the correct answer, as there are cases when no entry in a KB corresponds to an entity mention.

3.4 Entity disambiguation

To disambiguate entities, we make use of a neural end-to-end model based on a Bidirectional Long Short Term Memory (BiLSTM) and different types of embeddings. Specifically, the architecture follows the original model proposed by Kolitsas et al. [31] and depicted in Fig. 3.

The reason for using this architecture is that it performs both entity linking and entity disambiguation. Using it is therefore simpler and less prone to the propagation of errors. Moreover, this neural architecture does not need complex feature engineering. Thus, it is easy to adapt to multiple languages other than English.

For recognizing all entity mentions in a document, as Kolitsas et al. we made use of an empirical probabilistic table entity−map, as described previously in Sect. 3.2.

Our end-to-end EL model starts by encoding every input token into dense representations. This is done by concatenating word and character embeddings which are then fed into a BiLSTM [53] network. The BiLSTM network projects the document's mentions into a shared dimensional space, which has the same size as embeddings generated for the entities. The entity embedding is a collection of fixed continuous entity representations generated using the approach described by Ganea and Hofmann [29], as described in Sect. 3.3.

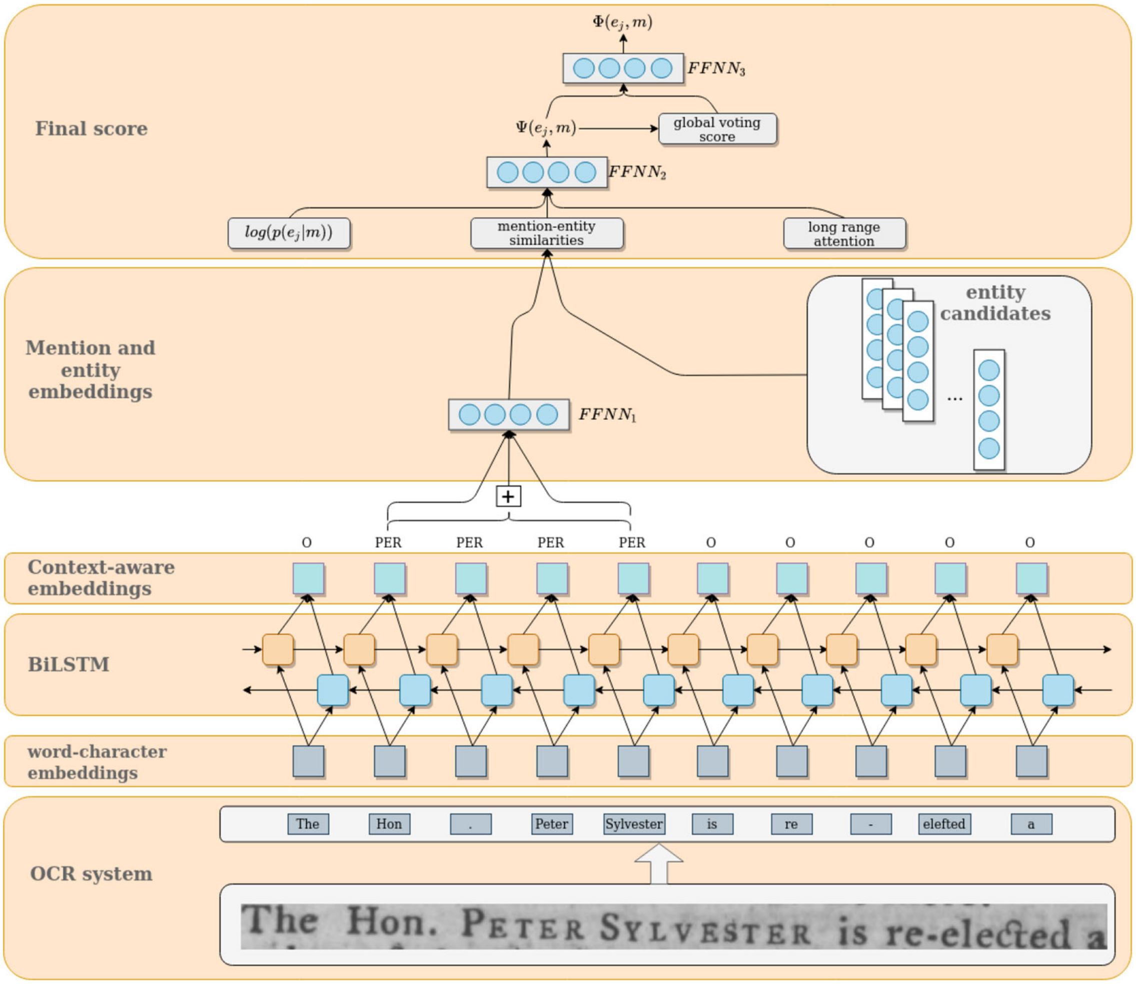


Fig. 3 Our global model architecture shown for the mention *Hon. Peter Sylvester* (from dev data of CLEF HIPE 2020 and published in [52]). The final score is used for both the mention linking and entity disambiguation decisions

To analyze long context dependencies of mentions, we use the attention mechanism defined by Ganea and Hofmann [29]. Specifically, this mechanism provides one context embedding per mention. This context is based on surrounding context words that are related to at least one of the candidate entities.

For each mention, the final score is determined by combining the log $p(e|m)$, similarity between a mention and a candidate entity, and the long-range context attention for this mention. Finally, the consistency between disambiguated entities within a document is promoted by a top layer in the neural network.

To minimize the impact of issues induced by historical data, we propose two techniques: a match correction that alleviates OCR-related issues (described in Sect. 3.5), and a method to add multilingual support, described in Sect. 3.6. Furthermore, we propose in Sect. 3.7 a post-processing filter to increase the performance of our EL system.

3.5 Match corrections

Multiple EL approaches [29,30,46], including the one used in this work, rely on the matching of entities and candidates using a probability table. If an entity is not listed in the probability table, the EL system cannot disambiguate it and, therefore, cannot propose candidates. In historical documents, the inability to match entities is a frequent problem, due to their inherent nature and processing, as explained in Sect. 1.

Multiple heuristics are used to analyze several surface name variations in order to increase the matching of entities in the probability table. These variations can deal with the casing (lower and upper capitalization), with the concatenation of surrounding words, the removal of stopwords, or the transliteration to Latin characters of some special characters like the accentuated ones.

Previous heuristics do not prevent missing matches. In that case, weighted Levenshtein distance is used to overcome more complex cases like transcription errors or spelling mistakes. We followed the idea exposed in [6] by using a mapping of OCR errors calculated on historical documents that helps in identifying common OCR mistakes (e.g. confusion between 'e' and 'c'). In this work, the average percentage mapping of OCR errors that are described in [6] is used to set up some weights in the Levenshtein distance.

3.6 Multilingualism

One of the biggest challenges in EL is the link of a mention for which a KB has no entry. This could happen either because it is known differently in specific languages [40,41] or because the KB is not large enough to cover the topic [45].

For instance, in Fig. 1h, we presented the case of an English document making reference to the Namur Gate using its French name, "Porte de Namur". While the English Wikipedia contains an entry regarding the Namur Gate, only in the French Wikipedia it is known as "Porte de Namur". This makes it impossible to find, on occasions, the correct entry to which a mention should be linked.

To solve this issue, we combine the probability tables generated by different languages, in order to create one multilingual probability table. In this way, the EL system can match the surface name of mentions with entries in multiple languages.

3.7 Filtering

To improve the accuracy of the non-NIL candidates provided by the EL system, we use a post-processing filter[14] based on heuristics and data provided by Wikidata and DBpedia. The goals of the filter are to: (1) Remove candidates which are unlikely such as disambiguation pages or people born after the document publication; (2) Verify that the tokens of a particular named entity are linked to the same candidates; (3) Fix redirection page issues; (4) Reorder the candidates based on their DBpedia type classification or how similar the candidate label is to the named entity to link.

The filter consists of four main steps, which are described as follows and presented graphically in Fig. 4.

The first step consists in querying Wikidata for five elements: *redirection_page* (boolean), *disambiguation_page* (boolean), *label* (string), *alternative_labels* (collection of strings), and *entry_year* (numeric). The first element helps us to find the correct page from which to extract the other elements.[15] For instance, the ID Q63832446 redirects automatically to Q4182026.[16] The second element, *disambiguation_page*, indicates whether we need to remove the candidate ID as a link cannot refer to an ambiguous entry, such as Moon (Q2432366)[17]. If the candidate ID is not a disambiguation page, we request from Wikidata the label (and the alternative labels, if any) associated with the entry in the language of analysis. For instance, the English entry of *Namur Gate* has as alternative label *Naamsepoort*.[18] Furthermore, we query Wikidata with the year in which the entry was conceived. For example, in the case of a person, it would be their birth year, while for a book (product), the year in which was published, or for a country (location), their inception date.

The second step filters the candidate ID based on their entry year and the publication year. If neither the entry nor the publication is associated with a year, this step is skipped. For fine-tuning, it is also possible to specify the types of mentions to be filtered by year.

The third step relies on querying DBpedia to determine whether the candidate ID exists in it and whether it is associated with specific categories defined for each mention type. If DBpedia does not contain the candidate ID or it does not link it to the specific categories, we request the same information to the different available DBpedia Chapters. We show in Table 2 the DBpedia types associated for each mention type. The categories associated with each mention type were manually defined. From this step, we generate three types of candidates:

- *Top* These IDs were considered in DBpedia to represent the mention type.
- *Middle* This type is for IDs for which it was impossible to retrieve information from either DBpedia or DBpedia Chapters. These IDs can be considered as part of bottom candidates depending on the filter configuration.
- *Bottom* It represents those candidates that were found in DBpedia, but whose DBpedia classification did not match the mention type.

Finally, the fourth step of the filter consists of sorting the three types of candidates defined in the previous step. The candidates are sorted based on incremental edit distances between the label (or alternative labels)[19] and the mention found in the text analyzed. The system breaks ties using the ordering in which the candidates were presented by the EL

[14] Code available at https://github.com/EMBEDDIA/NEL_Filter.

[15] Most of the redirections occur when two entries in Wikidata were merged. See: https://www.wikidata.org/wiki/Help:Redirects.

[16] https://www.wikidata.org/wiki/Q63832446.

[17] https://www.wikidata.org/wiki/Q2432366.

[18] https://www.wikidata.org/wiki/Q3399071.

[19] We take the string that produces the shortest distance.

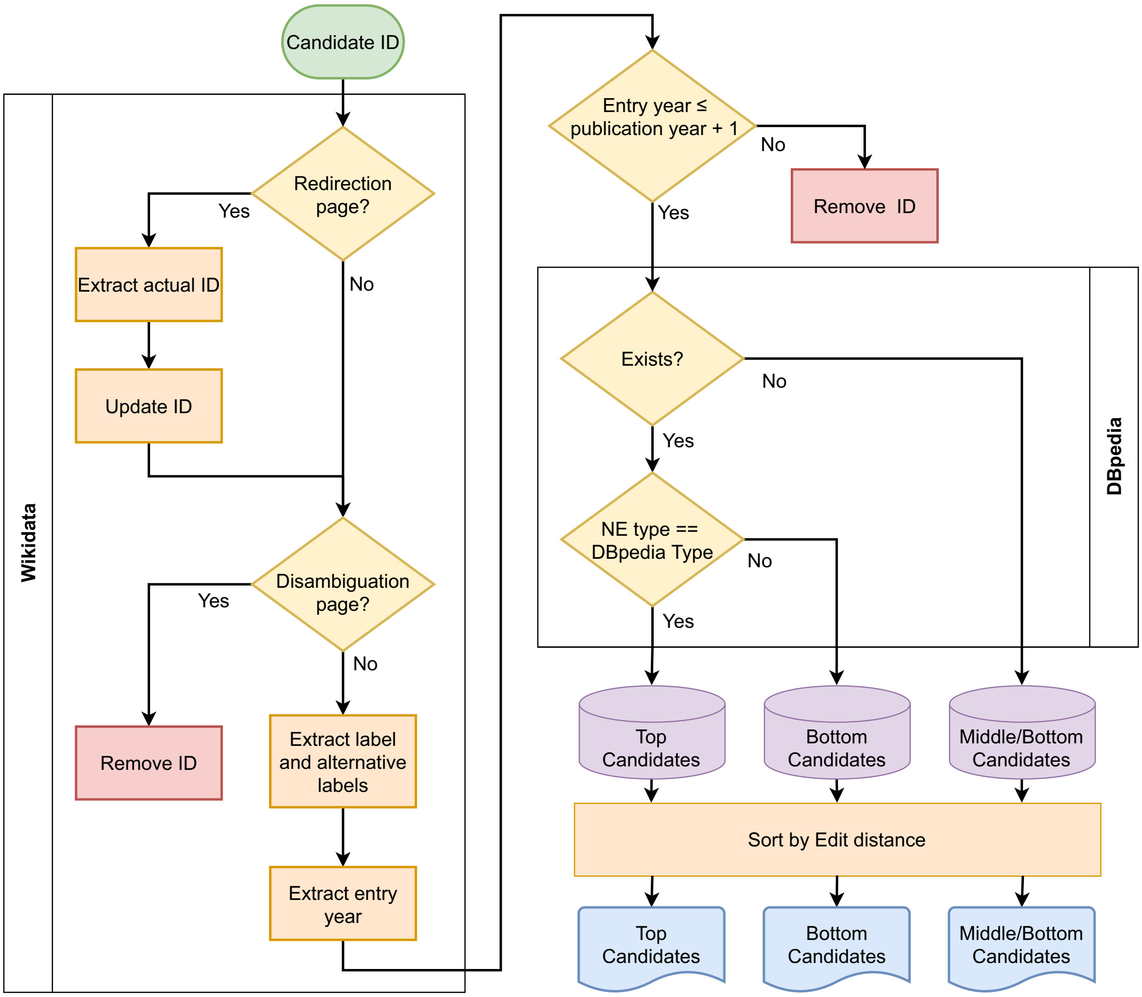


Fig. 4 Flow chart of the filtering module. This process is applied to all candidates provided by the EL system different to NIL

Table 2 Relation between each type of mentions, different to NIL, and their associated DBpedia types

Mention type	Associated DBpedia type
Location (LOC)	dbo:Location, dbo:Place, dbo:Settlement, dbo:Region, dbo:Building, dbo:Village, umbel-rc:Country, yago:YagoGeoEntity
Organization (ORG)	dbo:Organisation, umbel-rc:Business, dbc:Supraorganizations, yago:YagoGeoEntity
Person (PER)	foaf:Person, dbo:Person, dbo:Agent, dul:SocialPerson
Product (PRO)	dbo:Work, dbo:Newspaper, umbel-rc:Business, schema:CreativeWork, yago:TradeName106845599, yago:Product104007894

system. Once all the candidates have been sorted, they are printed as follows: 1) Top candidates 2) Middle candidates 3) NIL and 4) Bottom candidates. The addition of a NIL before the Bottom candidates stems from early experiments indicating that it is very unlikely that a mention would be linked to an ID that does not match the DBpedia classification. Therefore, we prefer to assign it the NIL link. This sorting can be of course easily adjusted.

We present in Fig. 5 an example of the filtering process for the named entity “Great Britain” found in a publication of 1868. As we can observe in Fig. 5, some of the candidate IDs refer to entities that started to exist long after the publication of the document. Also, not all of the proposed candidates are associated to the “location” type of DBpedia.

This filter architecture differs from the one proposed in our previous work [27], notably because the labels are obtained from Wikidata instead of DBpedia, and because we query for the DBpedia types to all the existing DBpedia services (i.e. including DBpedia and DBpedia Chapters), instead of just a subset of them. Also, we can filter by date different types of mentions and not only those related to people. Furthermore, we fix redirection pages and remove links of non-named enti-

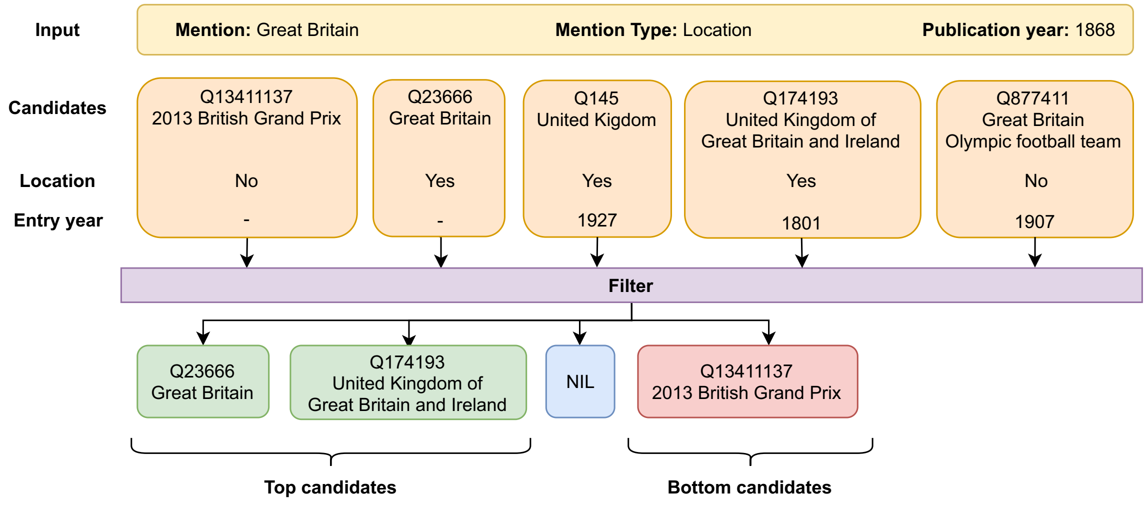


Fig. 5 Example of the filter application for the mention *Great Britain* in a 1868 publication. Before the filter, the candidates are arranged according to the EL system output

ties, which were previously ignored, and we sort middle and bottom candidates according to their edit distances.

4 Historical datasets

While EL on contemporary datasets can take advantage of an abundance of resources and tools [22,25,26,29–33], digitized and historical documents lack annotated resources [5,20,22,25,26,42]. Moreover, contemporary datasets and resources are generally not suitable for building accurate systems to be applied to historical datasets due to several issues, i.e. the variations in orthographic and grammatical rules, word variations, and also the fact that names of persons, organizations, or places could have significantly changed over time [5,20].

To the best of our knowledge, there are few publicly available corpora in the literature with entities manually annotated in historical documents [9,10,13]. Most of the EL datasets use contemporary documents [29–31] lacking the distinctive features found in historical documents.

In this paper, we focus on two datasets that contain historical documents in English, Finnish, French, German, and Swedish.

The first corpus was produced for the CLEF HIPE 2020 challenge[20] [54]. This corpus is composed of articles published between 1738 and 2019 in Swiss, Luxembourgish, and American newspapers. To build the corpus, the organizers randomly sampled articles from different newspapers according to predefined decades. For each newspaper, articles were randomly sampled among the first years of a set of predefined decades covering the lifespan of the newspaper, with the constraints to both have a title and more than 50 characters in length. The dataset was manually annotated by native speakers according to HIPE annotation guidelines [54].

The second corpus is the NewsEye dataset[21] [28,55] which is composed of a collection of annotated historical newspapers in French, German, Finnish, and Swedish. These newspapers were collected by the national libraries of France[22] (BnF), with documents from 1854 to 1946, Austria[23] (ONB) with documents from 1864 to 1933, and Finland[24] (NLF), with Finnish documents from 1852 and Swedish documents from 1848 to 1918.

Tables 3 and 4 describe the number of mentions by time period for the CLEF HIPE 2020 and the NewsEye datasets, respectively. The named entities from both datasets are classified according to their type and, when possible, linked to their Wikidata ID. The entities that do not exist in the Wikidata KB are linked to NIL entries.

5 Experimental settings

For all the languages, we utilize the multilingual pre-trained model MUSE[25]. Specifically, it is used for the entity embeddings and disambiguation model. The MUSE word embeddings are 300-sized, while the character embeddings are 50-sized.

[20] https://impresso.github.io/CLEF-HIPE-2020/.

[21] https://zenodo.org/record/4573313#.YH79nnUzY5k.

[22] https://www.bnf.fr.

[23] https://www.onb.ac.at.

[24] https://www.kansalliskirjasto.fi.

[25] https://github.com/facebookresearch/MUSE.

Table 3 Number of mentions by period of time in the CLEF HIPE dataset

Splits	German				English					French					
	1750–1800	1800–1850	1850–1900	1900–1950	1750–1800	1800–1850	1850–1900	1900–1950	1950–2000	1750–1800	1800–1850	1850–1900	1900–1950	1950–2000	> 2000
Train															
ORG	8	56	70	74	– –	– –	– –	– –	– –	10	38	50	116	88	14
LOC	12	84	105	111	– –	– –	– –	– –	– –	15	57	75	174	132	21
PERS	16	112	140	148	– –	– –	– –	– –	– –	20	76	100	232	176	28
PROD	20	140	175	185	– –	– –	– –	– –	– –	25	95	125	290	220	35
TIME	24	168	210	222	– –	– –	– –	– –	– –	30	114	150	348	264	42
Dev															
ORG	6	26	22	26	10	54	26	44	26	4	14	8	32	24	4
LOC	9	39	33	39	15	81	39	66	39	6	21	12	48	36	6
PERS	12	52	44	52	20	108	52	88	52	8	28	16	64	48	8
PROD	15	65	55	65	25	135	65	110	65	10	35	20	80	60	10
TIME	18	78	66	78	30	162	78	132	78	12	42	24	96	72	12
Test															
ORG	2	20	34	42	6	32	14	30	10	6	16	16	26	16	6
LOC	3	30	51	63	9	48	21	45	15	9	24	24	39	24	9
PERS	4	40	68	84	12	64	28	60	20	12	32	32	52	32	12
PROD	5	50	85	105	15	80	35	75	25	15	40	40	65	40	15
TIME	6	60	102	126	18	96	42	90	30	18	48	48	78	48	18

Table 4 Number of mentions by period of time in the NewsEye dataset

Splits	German		French			Finnish		Swedish		
	1850 –– 1900	1900 –– 1950	1800 –– 1850	1850 –– 1900	1900 –– 1950	1850 –– 1900	1900 –– 1950	1800 –– 1850	1850 –– 1900	1900 –– 1950
Train										
ORG	539	2571	169	100	1016	55	204	3	92	58
LOC	1437	3707	610	515	2930	401	578	13	620	352
PERS	1024	2082	920	299	3, 664	231	551	14	559	265
PROD	––	37	72	39	89	57	69	8	117	39
Dev										
ORG	9	114	18	45	71	11	26	1	11	5
LOC	72	191	64	45	226	22	75	8	68	72
PERS	31	118	64	24	187	11	66	4	59	21
PROD	––	4	2	6	3	2	10	2	2	13
Test										
ORG	21	116	24	9	184	15	6	––	11	3
LOC	157	340	161	67	369	42	42	8	90	42
PERS	122	123	155	36	272	51	40	21	87	34
PROD	1	2	6	9	6	3	4	1	11	3

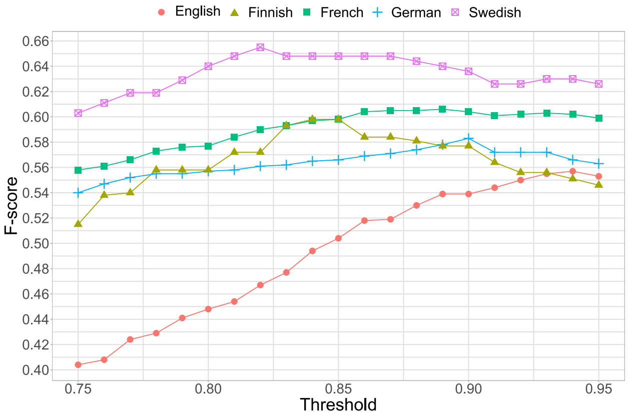


Fig. 6 F-score for different text distance thresholds to match mentions with OCR errors

As CLEF HIPE 2020 does not provide a training dataset for English, we make use of the contemporary corpus AIDA [37] for training purposes. Then the generated model is validated on the CLEF HIPE 2020 corpus.

Based on the statistical analysis of the training data, we defined the weighted Levenshtein distance ratio of 0.9, 0.94, 0.85, 0.89, and 0.82 for the languages German, English, Finnish, French, and Swedish, respectively, to search for other mentions in the probability table if the mention did not have a corresponding entry in the probability table (Fig. 6).

With respect to the post-processing filter, we query Wikidata, DBpedia, and DBpedia Chapters using their respective SPARQL Query Services.[26]. Ten DBpedia Chapters are used: Catalan, Basque, Greek, Indonesian, Dutch, French, German, Japanese, Korean, and Spanish. Furthermore, we explore two edit distance metrics: RapidFuzz Weight Ratio[27] and Weighted Levenshtein Distance[28] with specific costs defined by [6]. In addition, we explore whether candidates not found in DBpedia should be considered as middle candidates or bottom candidates. In total, we explore 18 different filters and their configuration is presented in Table 5.

For evaluating our methods, we compute their strict[29] F-score (F1) calculated for each language over the full corpus (micro-averaging).[30] Specifically, the F-score is defined as the harmonic mean between precision and recall, where precision is the fraction of correctly linked entity mentions that are generated by a system, and recall is the proportion of all entity mentions correctly linked over all the entity mentions that should be linked. We indicate that not all the mentions in the corpora have a corresponding entry in Wikidata. For instance, ambiguous names such as "Peter" or "Thomas", the gold standard is set to NIL: no link exists for the mention, and thus no link should be assigned to it.

[26] Wikidata: https://query.wikidata.org/, DBpedia: https://dbpedia.org/sparql, DBpedia Chapters: https://wiki.dbpedia.org/join/chapters.

[27] https://github.com/maxbachmann/rapidfuzz.

[28] https://pypi.org/project/weighted-levenshtein/.

[29] This means that a linked mention, in order to be counted as correct, must match both the gold standard's named entity boundary and its link.

[30] This was done using the tool at https://github.com/creat89/CLEF-HIPE-2020-scorer.

Table 5 Filter configurations used in this work

Filter	Edit distance	Mentions to filter by date	Middle candidates
1A	RapidFuzz W. Ratio	All	No
2A		None	
3A		Person	
4A		All	Yes
5A		None	
6A		Person	
1B	None	All	No
2B		None	
3B		Person	
4B		All	Yes
5B		None	
6B		Person	
1C	Weighted Levenshtein	All	No
2C		None	
3C		Person	
4C		All	Yes
5C		None	
6C		Person	

Motivated by the HIPE CLEF 2020 Shared Task [13], we also provided a more flexible evaluation by considering up to five answers for a mention (@5). In this case, an answer is considered correct if the referring Wikidata ID is among the top-5 candidates. Moreover, based on the work of [56], providing up to 5 candidates might produce the best user satisfaction in real applications.

6 Results

We present in Tables 6 and 7 the F-score obtained by each of the EL approaches detailed in Sect. 5, respectively, for the CLEF HIPE 2020 and NewsEye datasets. Tables 6 and 7 also contain the performance achieved by each post-processing filter applied to every base output generated by the EL systems.

From Tables 6 and 7 , we can notice that the match corrections, in general, improved the performance of the base EL candidates. Nonetheless, there are two languages and datasets, English CLEF HIPE 2020 and French NewsEye, where this approach reduced the performance of our EL systems. For the Swedish NewsEye dataset, only one configuration is negatively affected, i.e. when the match correction is coupled with a multilingual probability table.

Moreover, we can notice from Tables 6 and 7 that in four cases, CLEF HIPE 2020 English and NewsEye German, Finnish and French, the use of multilingual probability tables $p(e|m)$ reduced the performance of the EL systems. There are some partial exceptions, French CLEF HIPE 2020 and Swedish NewsEye, where multilingual probability tables $p(e|m)$ without match corrections performed better than monolingual probability tables without match corrections. Nevertheless, none of these two cases produces the best base EL performance in their respective languages.

As we can observe in Tables 6 and 7 , most of the EL configurations benefited from the application of a post-processing filter. The only exception is German CLEF HIPE 2020, where we used a monolingual probability table $p(e|m)$ and applied a match correction.

Although it is hard to observe in the first instance which filter performs best, we can notice certain patterns. In general, filters based on RapidFuzz Weight Ratio (filters A) generate the greatest number of top performances, especially in CLEF HIPE 2020 languages. Filters without re-ordering candidates based on edit distance (filters B), seem to generate the best performances for NewsEye Finnish. Finally, filters based on a Weighted Levenshtein distance (filters C) produce the fewer number of best performances in both corpora, i.e. in CLEF HIPE 2020 German and NewsEye Finnish.

We can observe as well in Tables 6 and 7 that, regardless of the corpus, mentions of type person are prone to be linked to entries that correspond to people born after the publication of the newspaper article. This can be clearly observed as filters 3 and 6 (which filter entities of type person) but also filters 1 and 4 (which filter all types of entities) produce the best performance.

In addition, we can notice in Tables 6 and 7 that for most datasets it is better not to use middle candidates (filters 1-3). The only exception is CLEF HIPE 2020 French, where it is

Table 6 Analysis of the performance (F-score) of our EL approach with different hyperparameters on the CLEF HIPE 2020 dataset

Match Cor.	Base	Filter																	
		1A	2A	3A	4A	5A	6A	1B	2B	3B	4B	5B	6B	1C	2C	3C	4C	5C	6C
German																			
$p(e\|m)$ =*multi*																			
False	0.517	0.527	0.536	0.539	0.527	0.535	0.538	0.522	0.528	0.531	0.521	0.527	0.530	0.531	0.540	**0.542**	0.530	0.539	0.541
True	0.555	0.561	0.571	0.574	0.561	0.570	0.573	0.557	0.565	0.567	0.556	0.564	0.566	0.566	0.577	***0.580***	0.565	0.576	0.579
$p(e\|m)$ =*mono*																			
False	0.534	0.529	0.533	**0.536**	0.527	0.531	0.534	0.529	0.532	0.535	0.527	0.530	0.534	0.528	0.532	0.535	0.527	0.530	0.534
True	**0.573**	0.562	0.567	0.570	0.561	0.565	0.568	0.562	0.566	0.569	0.561	0.564	0.568	0.561	0.568	0.571	0.560	0.566	0.569
English																			
$p(e\|m)$ =*multi*																			
False	0.572	**0.608**	0.599	0.601	0.606	0.595	0.597	0.601	0.592	0.595	0.599	0.588	0.590	0.601	0.592	0.595	0.599	0.588	0.590
True	0.556	**0.612**	0.597	0.606	**0.612**	0.595	0.604	0.604	0.588	0.597	0.604	0.586	0.595	0.606	0.590	0.599	0.606	0.588	0.597
$p(e\|m)$ =*mono*																			
False	0.606	**0.624**	**0.624**	**0.624**	**0.624**	0.621	0.621	0.621	0.621	0.621	0.621	0.619	0.619	0.610	0.610	0.610	0.610	0.608	0.608
True	0.597	***0.637***	0.630	***0.637***	0.635	0.626	0.633	0.635	0.628	0.635	0.633	0.624	0.630	0.624	0.617	0.624	0.621	0.612	0.619
French																			
$p(e\|m)$ =*multi*																			
False	0.602	0.601	0.609	0.613	0.604	0.613	**0.616**	0.598	0.604	0.609	0.600	0.608	0.612	0.597	0.603	0.608	0.599	0.606	0.611
True	0.625	0.630	0.636	0.642	0.634	0.638	***0.645***	0.624	0.629	0.635	0.628	0.631	0.638	0.623	0.627	0.634	0.627	0.629	0.637
$p(e\|m)$ =*mono*																			
False	0.594	0.594	0.602	0.605	0.597	0.604	**0.608**	0.591	0.598	0.601	0.593	0.600	0.604	0.591	0.598	0.601	0.593	0.600	0.604
True	0.630	0.631	0.637	0.641	0.634	0.639	**0.644**	0.625	0.631	0.636	0.629	0.633	0.639	0.627	0.633	0.637	0.631	0.635	0.641

Bold means the best performance on each configuration. Bold and italics mean best performance on each language

Table 7 Analysis of the performance (F-score) of our EL approach with different hyperparameters on the NewsEye dataset

Match Cor.	Base	Filter																	
		1A	2A	3A	4A	5A	6A	1B	2B	3B	4B	5B	6B	1C	2C	3C	4C	5C	6C
German																			
$p(e\|m)$ =*multi*																			
False	0.529	**0.607**	0.594	0.605	0.605	0.592	0.602	0.605	0.591	0.601	0.602	0.589	0.599	0.605	0.593	0.604	0.602	0.591	0.601
True	0.534	***0.620***	0.602	0.613	0.613	0.596	0.606	0.614	0.600	0.610	0.607	0.593	0.604	0.617	0.601	0.612	0.610	0.594	0.605
$p(e\|m)$ =*mono*																			
False	0.542	0.609	0.598	0.607	0.607	0.594	0.605	**0.610**	0.598	0.607	0.608	0.594	0.605	0.606	0.596	0.605	0.604	0.592	0.602
True	0.547	***0.620***	0.604	0.613	0.613	0.597	0.606	0.617	0.605	0.614	0.610	0.598	0.607	0.616	0.601	0.610	0.609	0.594	0.604
Finnish																			
$p(e\|m)$ =*multi*																			
False	0.595	0.618	0.598	0.608	0.623	0.603	0.613	0.623	0.603	0.613	**0.627**	0.608	0.618	0.623	0.598	0.608	**0.627**	0.603	0.613
True	0.615	0.652	0.642	0.652	0.652	0.642	0.652	0.652	0.642	0.652	0.652	0.642	0.652	**0.657**	0.642	0.652	**0.657**	0.642	0.652
$p(e\|m)$ =*mono*																			
False	0.632	0.647	0.637	0.642	0.647	0.637	0.642	**0.657**	0.647	0.652	**0.657**	0.647	0.652	0.652	0.637	0.642	0.652	0.637	0.642
True	0.652	0.676	0.672	0.676	0.676	0.662	0.676	***0.681***	0.676	***0.681***	***0.681***	0.667	***0.681***	***0.681***	0.672	0.676	***0.681***	0.662	0.676
French																			
$p(e\|m)$ =*multi*																			
False	0.551	0.626	0.626	**0.630**	0.623	0.623	0.627	0.623	0.623	0.627	0.6203	0.620	0.625	0.619	0.620	0.623	0.616	0.617	0.621
True	0.526	0.622	0.621	**0.625**	0.611	0.609	0.614	0.619	0.619	0.623	0.608	0.606	0.612	0.615	0.616	0.619	0.604	0.603	0.608
$p(e\|m)$ =*mono*																			
False	0.565	0.638	0.637	0.641	0.632	0.633	0.637	0.639	0.638	***0.642***	0.633	0.634	0.637	0.635	0.635	0.638	0.630	0.632	0.634
True	0.542	0.639	0.637	**0.641**	0.623	0.623	0.627	0.639	0.637	**0.641**	0.623	0.623	0.627	0.636	0.635	0.638	0.620	0.622	0.625
Swedish																			
$p(e\|m)$ =*multi*																			
False	0.590	0.617	0.614	0.627	0.611	0.614	0.621	0.621	0.617	**0.630**	0.614	0.617	0.624	0.617	0.614	0.627	0.611	0.614	0.621
True	0.580	0.643	0.630	0.653	0.637	0.621	0.646	0.646	0.633	***0.656***	0.640	0.624	0.650	0.643	0.630	0.653	0.637	0.621	0.646
$p(e\|m)$ =*mono*																			
False	0.577	0.598	0.601	**0.605**	0.592	0.595	0.598	0.598	0.601	**0.605**	0.592	0.595	0.598	0.598	0.601	**0.605**	0.592	0.595	0.598
True	0.599	0.637	0.637	**0.643**	0.630	0.621	0.637	0.637	0.637	**0.643**	0.630	0.621	0.637	0.637	0.637	**0.643**	0.630	0.621	0.637

Bold means the best performance on each configuration. Bold and italics mean best performance on each language

better to separate mentions not found in DBpedia as middle candidates (filters 4-6). For NewsEye Finnish, it seems that the EL system does not produce middle candidates, as it is equally good to use filters 1 and 3, or filters 4 and 6.

In Tables 8 and 9 , we present the performance of the EL systems calculating the F-score@5. As we can observe in Tables 8 and 9 , the increment in the performance for the base EL system when evaluating @1 and @5, indicates that in multiple cases the correct entry for a mention is found among the top-5 candidates.

Moreover, we can notice in Tables 8 and 9 that by applying a post-processing filter, we can still increase the performance. For instance, in NewsEye French we can observe an increase of up to 39.78%. By measuring the F-score@5, it is easier to observe certain patterns among the filters, such as the fact that filtering all mentions by date tends to be worse than only filtering by date for mentions of type person.

Based on the results presented in Tables 6, 7, 8, and 9 , we consider that the best performing configuration is a monolingual probability table with match correction and the filter 3A for all languages except Swedish. For this language, it is better to use a multilingual probability table with match correction and filter 3A. However, in order to take into account the specificities of each use case, a more sophisticated combination of various configurations should be explored.

7 Discussion

In this section, we present an analysis with respect to the probability tables used by the EL systems as well discussion regarding the obtained results.

7.1 Probability tables

This analysis is based on the gold standard data, specifically on the mentions that are linked to a Wikidata entry, i.e. no NILs. The goal is to improve the understanding of the results and the limitations of the proposed methods.

We start the analysis by introducing in Table 10 the number of mentions in the explored corpora that exists in each language and that are found in their respective Wikipedia KBs. As it can be observed in Table 10, for all the test splits, except for Swedish, at least 90% (91% − 96%) of the mentions are associated with an entry in the KBs. In comparison, for the Swedish test dataset, only 84% of mentions have a corresponding entry in the KBs. This means that not all the mentions, in a specific language, have a corresponding article in Wikipedia in the language of analysis. For instance, the entity “Porte de Namur” contains a corresponding entry in the Wikidata but not in the Finnish, German, or Swedish Wikipedia KBs. The consequence of this aspect is that, by default, monolingual probability tables $p(e|m)$ will not contain all entries necessary to link every mention.

The information presented in Table 10 is as well of relevance because, unlike recent works, such as [29,31], we analyze all the mentions even if they do not exist in a KB. In other words, our EL system is unaware of entities without a corresponding entry in the KBs. This aspect makes the EL task harder to perform, but more realistic, as in many cases, such as the CLEF HIPE 2020 Challenge, it is impossible to know beforehand the entities that will occur. Systems that analyze only mentions found in KBs tend to get better results. Nonetheless, the reason is that these systems reduce the pool of mentions to link and know a priori that these mentions will have a correct match in the KBs.

We present in Table 11, the number of mentions that match their surface form, either exactly or after applying a correction, with an entry in our probability tables $p(e|m)$ (described in Sect. 3.2). As it can be seen in Table 11, matching entities without applying match corrections (c.f. Sect. 7.3) is quite challenging. For some languages, such as Finnish and Swedish, less than 50% of the mentions match exactly with an entry in the probability tables $p(e|m)$. This shows that for these languages there is great variability and complexity on mentions’ surface forms, either due to aspects such as inflection and agglutination or to OCR errors. This phenomenon becomes significant on mentions of type person over the Finnish NewsEye dataset, where only 3% of the mentions can be matched in the probability tables.

We can notice in Table 11 that applying a match correction approach (c.f. Sect. 7.3) increases the number of entities that can be found in the probability tables. For instance, in NewsEye Finnish, the word “Berliiniin” (To Berlin) was spelled incorrectly as “Berliniin”, however, the match correction module found the correct entry in the KB, “Berliini” (Berlin). In some languages, like Finnish and Swedish, the matching increment is around 60%. Furthermore, we can increase the match of mentions of type person on the Finnish NewsEye dataset from 3% to 25%.

Finally, it is important to highlight that Table 11 allows us determining the maximum number of mentions that can be linked in a dataset if the disambiguation module would be perfect.

7.2 Multilingualism

Unlike previous recent literature [40,41], MELHISSA combined the probability tables generated by different languages, in order to create a unique multilingual probability table. While this solution clearly brought large performance improvements, several issues should be discussed. As presented in Table 11, the use of multilingual probability tables increased the number of mentions that match with an entry in the KBs. However, in multiple cases, the number of new men-

Table 8 Analysis of the performance, in terms of F-score@5, regarding our EL approach with different hyperparameters on the CLEF HIPE 2020 dataset

Match Cor.	Base	Filter																	
		1A	2A	3A	4A	5A	6A	1B	2B	3B	4B	5B	6B	1C	2C	3C	4C	5C	6C
German																			
$p(e\|m)$ =*multi*																			
False	0.595	0.606	0.621	0.621	0.606	0.621	0.621	0.606	0.620	0.620	0.606	0.620	0.620	0.606	**0.623**	**0.623**	0.606	**0.623**	**0.623**
True	0.639	0.651	0.669	0.669	0.651	0.669	0.669	0.651	0.668	0.668	0.651	0.668	0.668	0.651	***0.671***	***0.671***	0.651	***0.671***	***0.671***
$p(e\|m)$ =*mono*																			
False	0.586	0.596	**0.612**	**0.612**	0.596	**0.612**	**0.612**	0.596	**0.612**	**0.612**	0.596	**0.612**	**0.612**	0.596	**0.612**	**0.612**	0.596	**0.612**	**0.612**
True	0.629	0.639	**0.657**	**0.657**	0.639	**0.657**	**0.657**	0.639	**0.657**	**0.657**	0.639	**0.657**	**0.657**	0.639	**0.657**	**0.657**	0.639	**0.657**	**0.657**
English																			
$p(e\|m)$ =*multi*																			
False	0.623	**0.697**	**0.697**	**0.697**	**0.697**	**0.697**	**0.697**	**0.697**	**0.697**	**0.697**	**0.697**	**0.697**	**0.697**	**0.697**	**0.697**	**0.697**	**0.697**	**0.697**	**0.697**
True	0.612	**0.713**	**0.713**	**0.713**	**0.713**	**0.713**	**0.713**	**0.713**	**0.713**	**0.713**	**0.713**	**0.713**	**0.713**	**0.713**	**0.713**	**0.713**	**0.713**	**0.713**	**0.713**
$p(e\|m)$ =*mono*																			
False	0.637	**0.706**	**0.706**	**0.706**	**0.706**	**0.706**	**0.706**	**0.706**	**0.706**	**0.706**	**0.706**	**0.706**	**0.706**	**0.706**	**0.706**	**0.706**	**0.706**	**0.706**	**0.706**
True	0.635	***0.724***	***0.724***	***0.724***	***0.724***	***0.724***	***0.724***	***0.724***	***0.724***	***0.724***	***0.724***	***0.724***	***0.724***	***0.724***	***0.724***	***0.724***	***0.724***	***0.724***	***0.724***
French																			
$p(e\|m)$ =*multi*																			
False	0.650	0.675	0.688	0.688	0.675	0.688	0.688	0.674	**0.689**	**0.689**	0.674	**0.689**	**0.689**	0.674	0.688	0.688	0.674	0.688	0.688
True	0.680	0.717	0.731	0.731	0.717	0.731	0.731	0.716	**0.732**	**0.732**	0.716	**0.732**	**0.732**	0.716	0.731	0.731	0.716	0.731	0.731
$p(e\|m)$ =*mono*																			
False	0.646	0.667	0.679	0.680	0.667	0.680	0.680	0.667	0.680	**0.681**	0.667	**0.681**	**0.681**	0.667	0.679	0.679	0.667	0.677	0.677
True	0.688	0.719	0.731	0.733	0.719	0.733	0.733	0.719	0.732	***0.734***	0.719	***0.734***	***0.734***	0.719	0.732	0.732	0.719	0.731	0.731

Bold means the best performance on each configuration. Bold and italics mean best performance on each language

Table 9 Analysis of the performance, in terms of F-score@5, regarding our EL approach with different hyperparameters on the NewsEye dataset

Match Cor.	Base	Filter																	
		1A	2A	3A	4A	5A	6A	1B	2B	3B	4B	5B	6B	1C	2C	3C	4C	5C	6C
German																			
$p(e\|m)$ =*multi*																			
False	0.571	0.791	0.795	0.795	0.791	0.795	0.795	0.791	0.793	0.795	0.791	0.793	0.795	0.792	**0.796**	**0.796**	0.791	0.795	0.795
True	0.581	0.824	0.828	0.828	0.824	0.828	0.828	0.824	0.826	0.828	0.824	0.826	0.828	0.825	***0.829***	***0.829***	0.824	0.828	0.828
$p(e\|m)$ =*mono*																			
False	0.581	0.791	0.795	0.795	0.791	0.795	0.795	0.791	0.795	0.795	0.791	0.795	0.795	0.792	**0.796**	**0.796**	0.791	0.795	0.795
True	0.589	0.820	0.825	0.825	0.820	0.825	0.825	0.820	0.825	0.825	0.820	0.825	0.825	0.821	**0.826**	**0.826**	0.820	0.825	0.825
Finnish																			
$p(e\|m)$ =*multi*																			
False	0.634	**0.672**	**0.672**	**0.672**	**0.672**	**0.672**	**0.672**	**0.672**	**0.672**	**0.672**	**0.672**	**0.672**	**0.672**	**0.672**	**0.672**	**0.672**	**0.672**	**0.672**	**0.672**
True	0.659	0.706	***0.711***	***0.711***	0.706	***0.711***	***0.711***	0.706	***0.711***	***0.711***	0.706	***0.711***	***0.711***	0.706	***0.711***	***0.711***	0.706	***0.711***	***0.711***
$p(e\|m)$ =*mono*																			
False	0.647	**0.672**	**0.672**	**0.672**	**0.672**	**0.672**	**0.672**	**0.672**	**0.672**	**0.672**	**0.672**	**0.672**	**0.672**	**0.672**	**0.672**	**0.672**	**0.672**	**0.672**	**0.672**
True	0.672	0.706	***0.711***	***0.711***	0.706	***0.711***	***0.711***	0.706	***0.711***	***0.711***	0.706	***0.711***	***0.711***	0.706	***0.711***	***0.711***	0.706	***0.711***	***0.711***
French																			
$p(e\|m)$ =*multi*																			
False	0.589	0.752	**0.759**	**0.759**	0.752	**0.759**	**0.759**	0.752	**0.759**	**0.759**	0.750	0.757	0.757	0.752	**0.759**	**0.759**	0.750	0.757	0.757
True	0.563	0.780	***0.787***	***0.787***	0.780	***0.787***	***0.787***	0.780	***0.787***	***0.787***	0.778	0.785	0.785	0.780	***0.787***	***0.787***	0.778	0.785	0.785
$p(e\|m)$ =*mono*																			
False	0.593	0.742	0.749	0.749	0.742	0.749	0.749	0.742	0.749	0.749	0.742	0.749	0.749	0.744	**0.751**	**0.751**	0.742	0.749	0.749
True	0.573	0.773	0.780	0.780	0.773	0.780	0.780	0.773	0.780	0.780	0.773	0.780	0.780	0.774	**0.781**	**0.781**	0.773	0.780	0.780
Swedish																			
$p(e\|m)$ =*multi*																			
False	0.641	0.685	**0.695**	**0.695**	0.685	**0.695**	**0.695**	0.685	**0.695**	**0.695**	0.685	**0.695**	**0.695**	0.685	**0.695**	**0.695**	0.685	**0.695**	**0.695**
True	0.631	0.727	***0.736***	***0.736***	0.727	***0.736***	***0.736***	0.727	***0.736***	***0.736***	0.727	***0.736***	***0.736***	0.727	***0.736***	***0.736***	0.727	***0.736***	***0.736***
$p(e\|m)$ =*mono*																			
False	0.625	0.666	0.675	0.675	0.666	0.675	0.675	0.669	**0.678**	**0.678**	0.669	**0.678**	**0.678**	0.669	**0.678**	**0.678**	0.669	**0.678**	**0.678**
True	0.647	0.711	0.720	0.720	0.711	0.720	0.720	0.714	**0.723**	**0.723**	0.714	**0.723**	**0.723**	0.714	**0.723**	**0.723**	0.714	**0.723**	**0.723**

Bold means the best performance on each configuration. Bold and italics mean best performance on each language

Table 10 Amount of mentions of CLEF HIPE 2020 and NewsEye datasets that contain a corresponding entry in their language version of Wikipedia KB

Splits	CLEF HIPE 2020						NewsEye							
	English		French		German		Finnish		French		German		Swedish	
	Total	KB	Total	KB	Total	KB	Total	KB	Total	KB	Total	KB	Total	KB
Train														
ORG	−−	−−	554	509	247	208	92	81	352	291	165	149	72	50
LOC	−−	−−	3, 333	3, 244	2, 009	1, 922	837	801	1, 786	1, 743	454	427	834	792
PERS	−−	−−	1, 343	1, 102	849	667	270	215	1, 057	997	73	53	395	312
PROD	−−	−−	149	128	77	58	82	69	108	96	3	3	146	87
Total	−−	−−	5, 406	5, 008	3, 209	2, 868	1, 281	1, 166	3, 303	3, 127	695	632	1, 447	1, 241
Dev														
ORG	93	87	120	105	112	94	13	11	57	46	88	84	7	6
LOC	332	324	851	816	711	684	80	78	283	276	223	203	132	122
PERS	107	104	434	370	294	247	31	26	163	153	96	83	49	45
PROD	16	16	37	36	35	30	9	7	7	6	2	2	11	7
Total	549	532	1, 450	1, 335	1, 157	1, 058	133	122	510	481	409	372	199	180
Test														
ORG	45	40	105	104	99	92	2	1	79	69	61	54	5	2
LOC	184	182	926	898	696	685	78	75	481	466	269	255	121	113
PERS	46	43	271	238	188	142	31	31	208	199	73	58	56	42
PROD	8	8	43	40	42	35	5	5	15	15	3	3	13	8
Total	283	273	1, 345	1, 280	1, 025	954	116	112	783	749	406	370	195	165

tions matched is relatively low, with few exceptions within the testing splits for CLEF HIPE 2020 English and NewsEye French and Swedish. Furthermore, the increment of mention matches is relatively small in comparison to the number of entries added by merging the probability tables in different languages.

The increment on the matches contrasts with the reduction of the performance of the EL systems, in some cases, as shown in Tables 6 and 7 . Based on manual analysis, we determined three of the causes of this discrepancy.

First, the merge of the probability tables increases the number of possible candidates for each mention, which as a consequence requires a more robust EL method that can deal with the great number of candidates and their possible ambiguity.

Second, the fact that a mention and an entry match, at testing time, according to their surface name, does not ensure the location of a correct link. For instance, certain mentions, such as acronyms, can have different meanings in different languages. Therefore, the EL system might choose the incorrect entry, as it happened with the acronym “UE” that matches “Union Européenne” (European Union) in the French probability table but “University of the East” in the English one.

Third, and due to the nature of historical documents, OCR mistakes along with multilingual probability tables, can increase the ambiguity of entries for a determined mention. For instance, in CLEF HIPE 2020 English, the word France was detected by the OCR as “Fiance”[31]. This caused the EL system using a monolingual probability table to propose a NIL. However, the EL system using a multilingual probability table proposed as candidates “Georges P. Putnam” (Q5543134) and “Engagement” (Q157512).

7.3 Match correction

The use of match correction has proved to improve the performance of our EL systems as presented in Tables 6 and 7 . The main reason is that it increases the coverage of the mentions in the probability tables as seen in Table 11. In other words, aspects such as lexical variations, e.g. affixes and inflections, can be measured in order to find the best matching entry.

Furthermore, mentions with OCR errors can be more easily linked with their respective entry in the KBs. However, the application of a match correction can also have negative side effects. Similar to multilingual probability tables, match correction increases the number of entries to disambiguate. Consequently, some mentions might be matched to an incorrect entry.

This outcome agrees with other observations found in the literature, such as in [45], where the authors indicate that using an edit distance metric improves the matching of entities in noisy text.

7.4 Filtering

There are five reasons why, in most cases, the post-processing filters improved the performance of the EL systems.

First, the filter fixes redirection pages and removes disambiguation pages. Although both issues are infrequent, their fix can make a difference as to whether the actual best entry is positioned at the top or not.

Second, we use the filters to remove links to non-named entities. For example, in NewsEye German the token “Gast” (guest), a non-named entity, was tagged with a link to “Ausgasen” (Q778653). Also, the filters verify and fix that the same links are proposed to all the tokens of the analyzed named entity. For instance, in CLEF HIPE 2020 French, the named entity “New York” was once linked to Q60 and Q975653 for the token “New” but only to Q975653 for the token “York”. Without this fix, the EL system is penalized for either linking a non-existing named entity or splitting a named entity into multiple ones by proposing different candidates to multiple tokens.

Third, adding a NIL before the bottom candidates is a good technique to find mentions that do not have an entry in Wikipedia. However, its effect might not be visible unless we consider more than one candidate during the evaluation. Specifically, the effect of NIL can be seen in Tables 8 and 9 , where we can notice that for some languages, such as CLEF HIPE 2020 English, applying any of the filters resulted in the same score. The only common aspect between all the filters was the addition of a NIL before the bottom candidates. The addition of NIL was, in many cases, a contribution to the performance improvement when evaluating F-score@5. The results in Tables 8 and 9 show us as well that the base EL systems have a preference to link most mentions to an entry, rather than proposing a NIL.

Fourth, placing at the bottom candidates that do not match the mention type according to DBpedia is a good method to improve the performance of the EL system. This can be seen in the fact that positioning candidates not found in DBpedia at the bottom worked better than setting them in the middle.

Fifth, the use of supplementary information from additional KBs, such as DBpedia, has proved to be beneficial for the EL task before. For instance, Munelly et al. [42] use vCard[32] to find honorifics of people, while Brando et al. [24] use DBpedia and the BnF ontology[33] to retrieve the gender, honorifcs, family and given names of authors. In both cases, the supplementary information improved the matching of people. In our case, we use specific fields, from Wikidata

[31] HIPE-data-v1.3-test-en.tsv#L3070.

[32] https://www.w3.org/TR/vcard-rdf/.

[33] https://data.bnf.fr/.

Table 11 Amount of mentions that match their surface form with an entry existing in the probability tables $p(e|m)$

(a) CLEF HIPE 2020 dataset

Splits	English					French					German				
	Mentions	No cor.		Match Correction		Mentions	No cor.		Match Correction		Mentions	No cor.		Match Correction	
		Mono	Multi	Mono	Multi		Mono	Multi	Mono	Multi		Mono	Multi	Mono	Multi
Train															
ORG	–	–	–	–	–	554	252	266	465	473	247	82	83	172	173
LOC	–	–	–	–	–	3333	2221	2263	3022	3063	2009	1312	1338	1738	1767
PERS	–	–	–	–	–	1343	355	429	734	829	849	161	187	317	363
PROD	–	–	–	–	–	149	73	77	112	118	77	23	26	51	56
Total	–	–	–	–	–	5406	2904	3038	4348	4498	3209	1576	1632	2278	2,359
Dev															
ORG	93	44	45	72	73	120	46	47	97	98	112	19	19	72	72
LOC	332	210	211	272	272	851	595	617	759	767	711	494	504	613	625
PERS	107	50	52	61	62	434	140	158	242	265	294	87	98	131	143
PROD	16	9	9	12	12	37	13	17	30	30	35	13	15	29	31
Total	549	313	317	418	420	1450	795	840	1135	1167	1157	613	636	845	871
Test															
ORG	45	20	20	28	28	105	56	58	90	90	99	37	37	72	73
LOC	184	118	119	149	150	926	610	625	829	842	696	454	460	583	598
PERS	46	8	8	22	23	271	84	90	161	167	188	31	38	57	72
PROD	8	0	0	3	3	43	22	23	38	38	42	13	13	23	25
Total	283	146	147	202	204	1345	772	796	1118	1137	1025	535	548	735	768

Table 11 continued

(b) NewsEye dataset

Splits	Finnish					French					German					Swedish				
	Mentions	No cor.		Match Cor.		Mentions	No cor.		Match Cor.		Mentions	No cor.		Match Cor.		Mentions	No cor.		Match Cor.	
		Mono	Multi	Mono	Multi		Mono	Multi	Mono	Multi		Mono	Multi	Mono	Multi		Mono	Multi	Mono	Multi
Train																				
ORG	92	13	13	54	55	352	106	110	308	311	165	21	22	60	60	72	5	5	29	45
LOC	837	398	399	673	691	1786	1174	1201	1662	1681	454	321	321	383	386	834	485	516	722	756
PERS	270	21	35	78	110	1057	244	254	660	697	73	28	28	55	56	395	48	63	153	183
PROD	82	5	5	38	42	108	35	37	87	90	3	1	1	3	3	146	20	20	95	123
Total	1281	435	450	843	898	3303	1559	1602	2717	2779	695	371	372	501	505	1447	558	604	999	1107
Dev																				
ORG	13	3	3	7	7	57	22	22	51	52	88	46	46	69	69	7	1	1	4	5
LOC	80	37	38	62	62	283	209	211	262	263	223	175	175	209	210	132	81	81	120	123
PERS	31	2	2	13	13	163	59	62	116	122	96	27	30	54	56	49	14	24	24	37
PROD	9	0	1	2	4	7	5	5	6	6	2	2	2	2	2	11	1	1	10	10
Total	133	42	44	84	86	510	295	300	435	443	409	250	253	334	337	199	97	107	158	175
Test																				
ORG	2	0	0	1	1	79	33	36	65	69	61	14	14	49	49	5	1	1	2	2
LOC	78	47	47	65	65	481	304	310	442	444	269	184	184	230	231	121	75	81	108	113
PERS	31	1	2	7	8	208	87	91	135	140	73	22	22	52	56	56	14	16	27	29
PROD	5	0	0	1	4	15	6	6	14	14	3	0	0	3	3	13	2	2	11	12
Total	116	48	49	74	78	783	430	443	656	667	406	220	220	334	339	195	92	100	148	156

and DBpedia, like inception date and associated type, to filter the candidates proposed by our EL system.

Apart from the previous aspects, there are some particularities regarding the configuration of the filters that improved the performance of the EL systems. With respect to the edit distance metric, we observed that RapidFuzz Weight Ratio[34] produces in general the best ordering of candidates. The reason might be the fact that this edit-distance metric uses different heuristics, like alphabetically reordering the tokens or scaling the results based on the length of the strings.

There may also be other reasons why certain edit distances worked differently on specific datasets. For example, the Weighted Levenshtein might have worked better in English as it uses set weights to fix OCR errors found in English documents [8]. In addition, the implementation used only accepts ASCII characters, which might affect languages with diacritics such as French.

Although in Table 7, we observed that not using edit distance (B filters) performed better on NewsEye Finnish and Swedish, this outcome is caused by exactly two mentions, one in each language. Specifically, the label and/or alternative labels of the entries proposed by the EL systems caused to wrongly sort the top candidates. For instance, in NewsEye Swedish, the EL systems proposed for the mention "Ural" the entries "Uralfloden" (Q80240, Ural River) and "Uralbergen" (Q35600, Ural Mountains). While both entries do not match the mention's surface form, "Uralfloden"[35] has as an alternative label in Swedish the word "Ural", which produces an exact match. This makes the filter set on the first position "Uralfloden" instead of the correct entry "Uralbergen". In the case of Finnish, for the mention "Englannin" (England; in genitive singular form), the edit distances considered closer the entry "Englannin kuningaskunta" (Q179876, Kingdom of England) rather than "Englanti" (Q21, England). Based on the fact that only two mentions were affected by this aspect, we consider that, in real applications, it should always use an edit distance metric to reorder the candidates.

Regarding the filtering of entries by date, it is clear that it should always be done for mentions of type person. The reason is that most of the Wikipedia entries related to people contain a year of birth.

For the other types of mentions, i.e. location, organization, and product, the performance of the filter by date, seems to depend mostly on the dataset and how well the annotation was done or could be done.

For instance, we noticed that some locations were affected by the date filter due to ambiguities in the gold standard annotation. In Fig. 5, we presented the case of the mention "Great Britain" in a press article of 1868[36]. The gold standard annotation indicated that the correct entry is Q145, i.e. United Kingdom (of Great Britain and Northern Ireland). However, because the article was published in 1868, the correct entry should have been Q174193, i.e. United Kingdom of Great Britain and Ireland, which refers to the country that existed before the 1921 Anglo-Irish Treaty. The filter managed to propose the actual correct entry in second place, while removed the entry that matched the gold standard.

Some other annotation errors are due to the ambiguity of the entry in Wikidata or the impossibility of finding a better candidate. For example, in the French CLEF HIPE 2020, the mention "Val-de-Travers" in a 1798 document is associated in the gold standard to Q70526[37]. Nevertheless, despite the fact that the entry has for label "Val-de-Travers", it refers to a municipality created in 2009 (field "inception"). Thus, the filter removes it from the candidates. Nonetheless, in Wikipedia, it does not seem to exist a better candidate to annotate the entry. Some of the other entries, such as "Val-de-Travers District" or "Region of Val-de-Travers" make reference to relative modern locations too.

Although we found errors in the gold standard annotation, as indicated before, it should be mentioned that no adjustments or modifications were introduced in it. We highlighted them to improve the understanding of the obtained outcomes.

From a detailed analysis, we observed that most of the mention types benefited from the application of filters. The exception consisted in those belonging to organizations, in which the filter decreased the number of mentions with a correct entry positioned in the first place.

Nonetheless, when we evaluate the performance using F-score@5, this discrepancy is no longer observable. This means that the correct entry for organizations tends to be misplaced. The most probable reason is the small number of associated DBpedia types related to organizations as described in Table 2. This could also be related to a small coverage of organizations in DBpedia and DBpedia chapters.

8 Conclusions and future work

Historical documents are a window to the cultural and historical heritage of countries, regions, and languages. With their digitization, the accessibility of these documents has increased considerably, together with the need for information that can enrich these documents.

To enrich historical documents, digital humanities researchers have approached the natural language process-

[34] https://github.com/maxbachmann/rapidfuzz.

[35] https://www.wikidata.org/wiki/Q80240.

[36] It should be noted that the context surrounding the mention, indicates that "Great Britain" is referring to a country and not the island.

[37] https://www.wikidata.org/wiki/Q70526.

ing (NLP) community in order to have access to tools such as named entity recognition and entity linking. Although the use of NLP tools has expanded to multiple domains and types of documents, their use in historical corpora has been limited. Aspects such as optical character recognition (OCR) errors and spelling variations, which make NLP tasks harder to perform, have limited the number of tools available for historical documents.

In order to fill this gap, we presented MELHISSA, a Multilingual Entity Linking architecture for HIstorical preSS Articles. The main objective of this tool is to link mentions, such as names of people, organizations, and products, to entries in knowledge bases, such as Wikidata. Specifically, we created an end-to-end neural entity linking system that manages multiple languages and has been designed to surpass common errors found in historical documents.

The presented system was tested over two historical datasets, NewsEye and CLEF HIPE 2020, comprising five European languages: English, Finnish, French, German, and Swedish. We explored different configurations, such as the use of edit distances, multilingual probability tables, and post-processing filters, in order to create a reliable entity linking tool.

The obtained outcomes demonstrated that MELHISSA is a competitive tool that is able to get an F-score@1 of up to 0.681 and an F-score@5 of up to 0.787. We have observed that the use of multilingual probability tables can be useful in languages such as Swedish, while the use of a matching correction module can improve the pairing of mentions and entries in a knowledge base. Furthermore, the application of a post-processing filter can improve in general the entity linking performances in all languages.

To be precise, in MELHISSA, the use of multilingual probability tables allowed us to deal with entities that are either foreign words or found within a knowledge base in a different language than the one being analyzed. With respect to the matching correction module, it provided a way to match entities that had a different spelling, due to either OCR errors or language evolution. Finally, the use of the post-processing filter increased the performance of the MELHISSA thanks to its capacity of removing unlikely candidates, fixing issues from the knowledge bases and reordering the results.

Furthermore, although MELHISSA makes use of modern knowledge bases, i.e. Wikidata and DBpedia, which are based on contemporary sources, it is capable of linking historical entities. The main reason is that these knowledge bases represent, in many cases, not only current and contemporary entities but also historical ones. Nevertheless, we are aware that these knowledge bases lack coverage, either partial (certain languages) or total, regarding some historical entities due to their collaborative approach. Despite this shortcoming, we consider that MELHISSA's outcomes could point out missing historical entities and encourage experts to participate in the enrichment and/or improvement of Wikipedia, and in consequence of Wikidata and DBpedia.

In the future, there are multiple aspects that we would like to explore. In first place, we would like to extend the analysis of results using other evaluation tools, including statistical analysis, to determine which methods provide the best performance in real cases. In second place, we would like to explore new languages and contemporary documents to define whether specific configurations work better. This would allow us determining whether modules, such as matching correction and post-processing filters, can work for all languages and/or which aspects need to be improved to obtain a better generalization. In third place, we would like to see whether the use of diachronic embeddings could improve the entity matching. Entities might evolve over time, notably in their spelling and their meaning. Thus, the use of diachronic embeddings might increase the performance of an entity linking system. Besides, we consider that a comparative study of these embeddings on historical and contemporary documents could prove the effectiveness of MELHISSA steps (matching correction, post-processing filter) on larger time periods. Finally, we will promote MELHISSA to encourage researchers to use entity linking as a way to enrich the information available in historical documents.

Acknowledgements This work has been supported by the European Union's Horizon 2020 research and innovation programme under Grants 770299 (NewsEye) and 825153 (EMBEDDIA).

Author Contributions ELP helped in conceptualization, methodology, software, formal analysis, investigation, data curation, writing. LACD was involved in conceptualization, methodology, software, formal analysis, investigation, writing. JGM helped in conceptualization, methodology, project administration, validation, supervision, writing—review and editing. EB supervised the study and wrote the review and edited. AH contributed to formal analysis, investigation, data curation, visualization, writing—review and editing. AD helped in funding acquisition, conceptualization, methodology, project administration, validation, supervision, writing—review and editing. NS supervised and validated the study and wrote the review and edited. MC helped in supervision.

Funding Open Access funding provided by University of Helsinki including Helsinki University Central Hospital.

Data availability Our code is freely available at https://github.com/NewsEye/Named-Entity-Linking/tree/master/multilingual_entity_linking for the EL method and at https://github.com/EMBEDDIA/NEL_Filter for the filter module. The evaluation tool can be found at https://github.com/creat89/CLEF-HIPE-2020-scorer. The CLEF HIPE 2020 dataset is available at https://impresso.github.io/CLEF-HIPE-2020/ and the NewsEye dataset at https://zenodo.org/record/4573313#.YH79nnUzY5k.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Oberbichler, S., Pfanzelter, E., Marjanen, J., Hechl, S.: Doing historical research with digital newspapers: perspectives of dh scholars. EuropeanaTech Insight, 16: Newspapers (2020). https://pro.europeana.eu/page/issue-11-generous-interfaces
2. Bair, S., Carlson, S.: Where keywords fail: using metadata to facilitate digital humanities scholarship. J. Libr. Metadata **8**(3), 249–262 (2008)
3. Wevers, M., Koolen, M.: Digital begriffsgeschichte: tracing semantic change using word embeddings. Hist. Methods J. Quant. Interdiscip. His. **53**(4), 226–243 (2020)
4. Hechl, S., Langlais, P.C., Marjanen, J., Oberbichler, S., Pfanzelter, E.: Digital interfaces of historical newspapers: opportunities, restrictions and recommendations. J. Data Mining Digital, Hum (2021)
5. Linhares Pontes, E., Hamdi, A., Sidere, N., Doucet, A.: Impact of OCR quality on named entity linking. In: Digital libraries at the crossroads of digital information for the future - 21st international conference on Asia-Pacific digital libraries, ICADL 2019, Kuala Lumpur, Malaysia, November 4-7, 2019, Proceedings, pp. 102–115 (2019). https://doi.org/10.1007/978-3-030-34058-2_11
6. Nguyen, T.T.H., Jatowt, A., Coustaty, M., Nguyen, N.V., Doucet, A.: Deep Statistical Analysis of OCR Errors for Effective Post-OCR Processing. In: Proceedings of the 18th joint conference on digital libraries, JCDL '19, p. 29–38. IEEE Press (2019). https://doi.org/10.1109/JCDL.2019.00015
7. Linhares Pontes, E., Moreno, J.G., Doucet, A.: Linking named entities across languages using multilingual word embeddings. In: Proceedings of the ACM/IEEE joint conference on digital libraries in 2020, JCDL '20, p. 329–332. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3383583.3398597
8. Nguyen, N.K., Boros, E., Lejeune, G., Doucet, A.: Impact analysis of document digitization on event extraction. In: 4th Workshop on natural language for artificial intelligence (NL4AI 2020) co-located with the 19th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2020), vol. 2735, pp. 17–28 (2020)
9. Boroş, E., Hamdi, A., Pontes, E.L., Cabrera-Diego, L.A., Moreno, J.G., Sidere, N., Doucet, A.: Alleviating digitization errors in named entity recognition for historical documents. In: Proceedings of the 24th conference on computational natural language learning, pp. 431–441 (2020)
10. Boros, E., Linhares Pontes, E., Cabrera-Diego, L.A., Hamdi, A., Moreno, J.G., Sidère, N., Doucet, A.: Robust named entity recognition and linking on historical multilingual documents. In: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (eds.) CLEF 2020 Working Notes. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. CEUR-WS (2020)
11. Boroş, E., Romero, V., Maarand, M., Zenklová, K., Křečková, J., Vidal, E., Stutzmann, D., Kermorvant, C.: A comparison of sequential and combined approaches for named entity recognition in a corpus of handwritten medieval charters. In: 2020 17th International conference on frontiers in handwriting recognition (ICFHR), pp. 79–84. IEEE (2020)
12. Oberbichler, S., Boroş, E., Doucet, A., Marjanen, J., Pfanzelter, E., Rautiainen, J., Toivonen, H., Tolonen, M.: Integrated interdisciplinary workflows for research on historical newspapers: perspectives from humanities scholars, computer scientists, and librarians. J. Assoc. Inf. Sci, Technol (2021)
13. Ehrmann, M., Romanello, M., Bircher, S., Clematide, S.: Introducing the CLEF 2020 HIPE shared task: Named entity recognition and linking on historical newspapers. In: J.M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M.J. Silva, F. Martins (eds.) Proceedings of the 42nd European conference on IR research (ECIR 2020), vol. 2, pp. 524–532. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-45442-5_68
14. Humbel, M., Nyhan, J., Vlachidis, A., Sloan, K., Ortolja-Baird, A.: Named-entity recognition for early modern textual documents: a review of capabilities and challenges with strategies for the future. J. Doc. (2021). https://doi.org/10.1108/JD-02-2021-0032
15. Nguyen, T.T.H., Jatowt, A., Coustaty, M., Doucet, A.: Survey of post-OCR processing approaches. ACM Comput. Surv. **54**(6), 1 (2021)
16. Rigaud, C., Doucet, A., Coustaty, M., Moreux, J.P.: ICDAR 2019 Competition on Post-OCR Text Correction. In: 2019 international conference on document analysis and recognition (ICDAR), pp. 1588–1593 (2019). https://doi.org/10.1109/ICDAR.2019.00255
17. Gefen, A.: Les enjeux épistémologiques des humanités numériques. Socio (2015). https://doi.org/10.4000/socio.1296
18. Chiron, G., Doucet, A., Coustaty, M., Visani, M., Moreux, J.P.: Impact of OCR errors on the use of digital libraries: towards a better access to information. In: Proceedings of the 17th ACM/IEEE joint conference on digital libraries, pp. 249–252. IEEE Press (2017)
19. Smith, D.A., Crane, G.: Disambiguating geographic names in a historical digital library. In: Proceedings of the 5th European conference on research and advanced technology for digital libraries, ECDL '01, p. 127–136. Springer-Verlag, Darmstadt, Germany (2001). https://doi.org/10.1007/3-540-44796-2_12
20. Heino, E., Tamper, M., Mäkelä, E., Leskinen, P., Ikkala, E., Tuominen, J., Koho, M., Hyvönen, E.: Named entity linking in a complex domain: Case second world war history. In: Gracia, J., Bond, F., McCrae, J.P., Buitelaar, P., Chiarcos, C., Hellmann, S. (eds.) Language, Data, and Knowledge, pp. 120–133. Springer, Galway, Ireland (2017). https://doi.org/10.1007/978-3-319-59888-8_10
21. Shen, W., Wang, J., Han, J.: Entity linking with a knowledge base: issues, techniques, and solutions. IEEE Trans. Knowl. Data Eng. **27**(2), 443–460 (2015). https://doi.org/10.1109/TKDE.2014.2327028
22. van Hooland, S., De Wilde, M., Verborgh, R., Steiner, T., Van de Walle, R.: Exploring entity recognition and disambiguation for cultural heritage collections. Digital Scholarship Hum. **30**(2), 262–279 (2013). https://doi.org/10.1093/llc/fqt067
23. Brando, C., Frontini, F., Ganascia, J.G.: Disambiguation of Named Entities in Cultural Heritage Texts Using Linked Data Sets. In: T. Morzy, P. Valduriez, L. Bellatreche (eds.) First international workshop on semantic web for cultural heritage, SW4CH 2015, Communications in computer and information science, vol. 539, pp. 505–514. Springer, Poitiers, France (2015). https://doi.org/10.1007/978-3-319-23201-0_51
24. Brando, C., Frontini, F., Ganascia, J.G.: REDEN: named entity linking in digital literary editions using linked data sets. Complex Syst. Inf. Model. Quarter. **2016**(7), 60–80 (2016). https://doi.org/10.7250/csimq.2016-7.04

25. Munnelly, G., Lawless, S.: Investigating entity linking in early english legal documents. In: Proceedings of the 18th ACM/IEEE on joint conference on digital libraries, JCDL '18, p. 59–68. Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3197026.3197055
26. Ruiz, P., Poibeau, T.: Mapping the bentham corpus: concept-based navigation. J. Data Mining Digital Humanities. **Special Issue: Digital Humanities between knowledge and know-how (Atelier Digit_Hum)** (2019). https://hal.archives-ouvertes.fr/hal-01915730
27. Linhares Pontes, E., Cabrera-Diego, L.A., Moreno, J.G., Boros, E., Hamdi, A., Sidère, N., Coustaty, M., Doucet, A.: Entity linking for historical documents: challenges and solutions. In: Ishita, E., Pang, N.L.S., Zhou, L. (eds.) Digital Libraries at Times of Massive Societal Transition, pp. 215–231. Springer, Cham (2020)
28. Hamdi, A., Boroş, E., Pontes, E.L., Nguyen, T.T.H., Hackl, G., Moreno, J.G., Doucet, A.: A multilingual dataset for named entity recognition, entity linking and stance detection in historical newspapers. In: Proceedings of the 44rd International ACM SIGIR conference on research and development in information retrieval (2021)
29. Ganea, O.E., Hofmann, T.: Deep joint entity disambiguation with local neural attention. In: Proceedings of the 2017 conference on empirical methods in natural language processing, pp. 2619–2629. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/D17-1277
30. Onoe, Y., Durrett, G.: Fine-grained entity typing for domain independent entity linking. Proc. AAAI Conf. Artif. Intell. **34**, 8576–8583 (2020)
31. Kolitsas, N., Ganea, O.E., Hofmann, T.: End-to-end neural entity linking. In: Proceedings of the 22nd conference on computational natural language learning, pp. 519–529. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/K18-1050
32. Cucerzan, S.: Large-scale named entity disambiguation based on Wikipedia data. In: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), pp. 708–716. Association for computational linguistics, Prague, Czech Republic (2007). https://www.aclweb.org/anthology/D07-1074
33. Broscheit, S.: Investigating entity knowledge in BERT with simple neural end-to-end entity linking. In: Proceedings of the 23rd conference on computational natural language learning (CoNLL), pp. 677–685. Association for computational linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/K19-1063. https://aclanthology.org/K19-1063
34. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In: Proceedings of the seventh conference on natural language learning at HLT-NAACL 2003, pp. 142–147 (2003). https://aclanthology.org/W03-0419
35. Chen, S., Wang, J., Jiang, F., Lin, C.Y.: Improving entity linking by modeling latent entity type information. In: Proceedings of the AAAI conference on artificial intelligence, **34**, 7529–7537 (2020)
36. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423
37. Hoffart, J., Yosef, M.A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G.: Robust disambiguation of named entities in text. In: Proceedings of the 2011 conference on empirical methods in natural language processing, pp. 782–792. Association for Computational Linguistics, Edinburgh, Scotland, UK. (2011). https://www.aclweb.org/anthology/D11-1072
38. Mosallam, Y., Abi-Haidar, A., Ganascia, J.G.: Unsupervised named entity recognition and disambiguation: an application to old French journals. In: Perner, P. (ed.) Advances in Data Mining: Applications and Theoretical Aspects, pp. 12–23. Springer, St. Petersburg, Russia (2014)
39. Rijhwani, S., Xie, J., Neubig, G., Carbonell, J.: Zero-shot neural transfer for cross-lingual entity linking. In: Thirty-Third AAAI conference on artificial intelligence (AAAI). Honolulu, Hawaii (2019). https://doi.org/10.1609/aaai.v33i01.33016924
40. Zhou, S., Rijhwani, S., Neubig, G.: Towards zero-resource cross-lingual entity linking. In: Proceedings of the 2nd workshop on deep learning approaches for low-resource NLP (DeepLo 2019), pp. 243–252. ACL, China (2019). https://doi.org/10.18653/v1/D19-6127
41. Zhou, S., Rijhwani, S., Wieting, J., Carbonell, J., Neubig, G.: Improving candidate generation for low-resource cross-lingual entity linking. Trans. Assoc. Comput. Linguist. **8**, 109–124 (2020)
42. Munnelly, G., Pandit, H.J., Lawless, S.: Exploring linked data for the automatic enrichment of historical archives. In: European Semantic Web Conference, pp. 423–433. Springer (2018). https://doi.org/10.1007/978-3-319-98192-5_57
43. Huet, T., Biega, J., Suchanek, F.M.: Mining history with le monde. In: Proceedings of the 2013 workshop on automated knowledge base construction, AKBC '13, p. 49–54. Association for Computing Machinery, New York, NY, USA (2013). https://doi.org/10.1145/2509558.2509567
44. Pellissier Tanon, T., Weikum, G., Suchanek, F.: YAGO 4: A reason-able knowledge base. In: A. Harth, S. Kirrane, A.C. Ngonga Ngomo, H. Paulheim, A. Rula, A.L. Gentile, P. Haase, M. Cochez (eds.) Proceedings of the 17th International conference, ESWC 2020, The Semantic Web, pp. 583–596. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-49461-2_34
45. Klie, J.C., Eckart de Castilho, R., Gurevych, I.: From zero to hero: human-in-the-loop entity linking in low resource domains. In: Proceedings of the 58th Annual meeting of the association for computational linguistics, pp. 6982–6993. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.624
46. Abramitzky, R., Mill, R., Pérez, S.: Linking individuals across historical sources: a fully automated approach. Hist. Methods J Quant. Interdiscip. Hist. **53**(2), 94–111 (2020)
47. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., Kleef, Pv., Auer, S., Bizer, C.: DBpedia: a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web J. **6**(2), 167–195 (2015). https://doi.org/10.3233/SW-140134
48. Moreno, J.G., Besançon, R., Beaumont, R., D'hondt, E., Ligozat, A.L., Rosset, S., Tannier, X., Grau, B.: Combining word and entity embeddings for entity linking. In: European Semantic Web Conference, pp. 337–352. Springer (2017)
49. Agirre, E., Barrena, A., de Lacalle, O.L., Soroa, A., Fernando, S., Stevenson, M.: Matching cultural heritage items to wikipedia. In: Eight International conference on language resources and evaluation (LREC) (2012)
50. Frontini, F., Brando, C., Ganascia, J.G.: Semantic web based named entity linking for digital humanities and heritage texts. In: Proceedings of the first international workshop semantic web for scientific heritage at the 12th ESWC 2015 Conference, vol. 1364 (2015)
51. De Wilde, M.: Improving retrieval of historical content with entity linking. In: Morzy, T., Valduriez, P., Bellatreche, L. (eds.) New Trends in Databases and Information Systems (ADBIS 2015), pp. 498–504. Springer, Berlin (2015)

52. Gazette of the United-States. (New York, New York, U.S.A). In: Chronicling America: Historic American Newspapers. Library of Congress (29-May-1790). https://chroniclingamerica.loc.gov/lccn/sn83030483/1790-05-29/ed-1/seq-3/. Accessed on April 2021
53. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
54. Ehrmann, R., Clematide, F.: HIPE: Shared Task Participation Guidelines (2020). https://doi.org/10.5281/zenodo.3677171
55. Doucet, A., Gasteiner, M., Granroth-Wilding, M., Kaiser, M., Kaukonen, M., Labahn, R., Moreux, J.P., Muehlberger, G., Pfanzelter, E., Thérenty, M.È., Toivonen, H., Tolonen, M.: NewsEye: A digital investigator for historical newspapers. In: 15th Annual international conference of the alliance of digital humanities organizations, DH 2020. Ottawa, Canada (2020)
56. Han, B., Shah, C., Saelid, D.: Users perception of search-engine biases and satisfaction. In: Boratto, L., Faralli, S., Marras, M., Stilo, G. (eds.) Advances in Bias and Fairness in Information Retrieval, pp. 14–24. Springer, Cham (2021)
57. Gazette of the United-States. (New York, New York, U.S.A). In: Chronicling America: Historic American Newspapers. Library of congress (02-Jan-1790). https://chroniclingamerica.loc.gov/lccn/sn83030483/1790-01-02/ed-1/seq-4/. Accessed on April 2021
58. Gazette of the United-States. (New York, New York, U.S.A). In: Chronicling America: Historic American Newspapers. Library of congress (03-Mar-1790). https://chroniclingamerica.loc.gov/lccn/sn83030483/1790-03-03/ed-1/seq-4/. Accessed on April 2021
59. Vossische Zeitung. (Berlin , Germany). Staatsbibliothek zu Berlin (11-Feb-1857). https://dfg-viewer.de/show/?set%5Bmets%5D=https://content.staatsbibliothek-berlin.de/zefys/SNP27112366-18570211-0-0-0-0.xml. Accessed on April 2021
60. CharitonCourier.(Keytesville,CharitonCounty,Missouri,U.S.A). In: Chronicling America: Historic American newspapers. Library of congress (13-Feb-1890). Accessed on April 2021
61. Le Liberateur du Sud-Ouest : organe rgional du Parti populaire francais. (Bordeaux , France). Bibliothque nationale de France (3-Dec-1936). https://gallica.bnf.fr/ark:/12148/bpt6k55631820. Accessed on April 2021
62. Les Affiches de Paris (Paris , France). Bibliothque nationale de France (31-Dec-1750). https://gallica.bnf.fr/ark:/12148/bpt6k10531388. Accessed on April 2021