Entity Linking for Historical Documents: Challenges and Solutions

Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Jose G. Moreno, Emanuela Boros, Elvys Linhares Pontes, Ahmed Hamdi, Nicolas Sidère, Mickaël Coustaty, Antoine Doucet

Entity Linking for Historical Documents: Challenges and Solutions

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

[1] University of La Rochelle, L3i, F-17000, La Rochelle, France
firstname.lastname@univ-lr.fr
[2] University of Toulouse, IRIT, UMR 5505 CNRS, F-31000, Toulouse, France
jose.moreno@irit.fr

Abstract. Named entities (NEs) are among the most relevant type of information that can be used to efficiently index and retrieve digital documents. Furthermore, the use of Entity Linking (EL) to disambiguate and relate NEs to knowledge bases, provides supplementary information which can be useful to differentiate ambiguous elements such as geographical locations and peoples' names. In historical documents, the detection and disambiguation of NEs is a challenge. Most historical documents are converted into plain text using an optical character recognition (OCR) system at the expense of some noise. Documents in digital libraries will, therefore, be indexed with errors that may hinder their accessibility. OCR errors affect not only document indexing but the detection, disambiguation, and linking of NEs. This paper aims at analysing the performance of different EL approaches on two multilingual historical corpora, CLEF HIPE 2020 (English, French, German) and NewsEye (Finnish, French, German, Swedish), while proposes several techniques for alleviating the impact of historical data problems on the EL task. Our findings indicate that the proposed approaches not only outperform the baseline in both corpora but additionally they considerably reduce the impact of historical document issues on different subjects and languages.

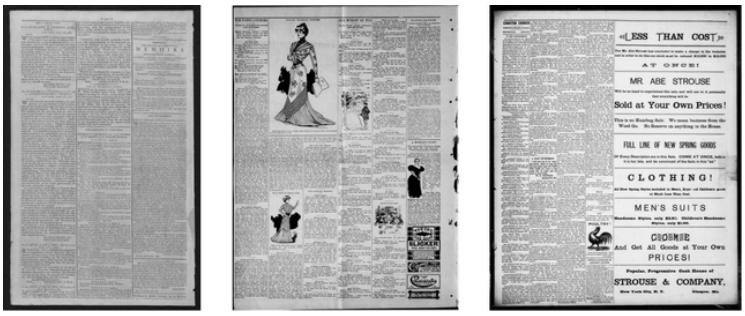
Keywords: Entity linking · Deep learning · Historical data · Digital libraries.

1 Introduction

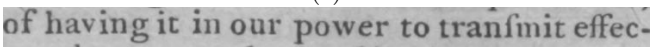
Historical documents are an essential resource in the understanding of our cultural heritage. The development of recent technologies, such as optical character recognition (OCR) systems, allows the digitisation of physical documents and the extraction of the textual content. Digitisation provides two major advantages in Digital Humanities: the exponential increase of target audiences, and the preservation of original documents from any damage when accessing them. The recent interest in massive digitisation raises multiple challenges to content providers including indexing, categorisation, searching, to mention a few. Although these

challenges also exist when dealing with contemporary text documents, digitised version augments each challenge because of inherent problems associated with the source quality (natural degradation of the documents) and to the digitisation process itself (e.g. image quality and OCR bias).

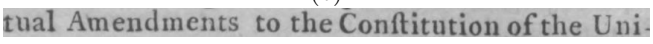
While the number of works in natural language processing (NLP) and information retrieval (IR) domains concerning contemporary documents has known an important raise during the last decade, it has not been the case for historical documents. One of the main reasons is the additional difficulties that NLP and IR systems have to face regarding historical documents. For instance, tools need to know how to deal correctly with errors produced by OCR systems. Moreover, historical languages may contain a number of spelling variations with respect to modern languages, that might be difficult to recognise, as orthographic conventions can be reformed from time to time. Finally, some historic documents may also contain cases where the name of places is in a language different to the main text one. These particularities have then a significant impact on NLP and IR applications over historical documents.



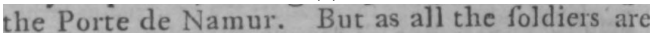
(a)



(b)



(c)



(d)

Fig. 1. Examples of historical documents from the Chronicling America newspapers used in CLEF HIPE 2020.

To illustrate some of the aforementioned problems, let us consider Figure 1(a) which includes some English documents used in the evaluation campaign CLEF HIPE 2020 [9]. Figure 1(b) and Figure 1(c) are zoomed and cropped portions of most left document presented in Figure 1(a). We can observe in these images a common characteristic found in multiple historical documents, the presence of a *Long S* ("*ſ*"), a character that is frequently confused by OCR systems for an

"l" or "f" given its geometrical similarity. Figure 1(b) illustrates a case where the word "tranΓmit" was recognised as "tranlinit" by a state-of-the-art OCR system.[3] Figure 1(c) illustrates a similar case where the word "ConΓtitution" was recognised as "Conftitution"[4] which makes harder for an automatic system to recognise that this document concerns the *Constitution of the Unites States of America*[5]. In Figure 1(d), we observe a case where an article uses the French name "Porte de Namur" to make reference to "Namur Gate".[6]

Apart from digitising and recognising the text, the processing of historical documents consists as well on extracting metadata from these documents. 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 documents' content [12]. These entities can represent aspects such as people, places, organisations, and events. Nonetheless, historical documents may contain duplicated and ambiguous information about named entities due to the heterogeneity and the mix of temporal references [30,13]. A disambiguation process is thus essential to distinguish named entities to be further utilised by search systems in digital libraries.

Entity linking (EL) aims to recognise, 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 [29]. Compared to contemporary data, few works in the state of the art have studied the EL task on historical documents [30,16,3,4,13,23,28] and OCR-processed documents [20].

In this paper, we present a deep learning EL approach to disambiguate entities on historical documents. We investigate the issues of historical documents and propose several techniques to overcome and reduce the impact of these issues in the EL task. Moreover, our EL approach decreases possible bias by not limiting or focusing the explored entities to a specific dataset. We evaluate our methods in two recent historical corpora, CLEF HIPE 2020 [9], and NewsEye datasets, that are composed of documents in English, Finnish, French, German, and Swedish. Our study shows that our techniques improve the performance of EL systems and partially solve the issues of historical data.

This paper is organised as follows: we describe and survey the EL task on historical data in Section 2. Next, the CLEF HIPE 2020 and NewsEye datasets are described in Section 3. We detail our multilingual approach in Section 4. Then the experiments and the results are discussed in Sections 5 and 6. Lastly, we provide the conclusion and some final comments in Section 7.

[3] HIPE-data-v1.3-test-masked-bundle5-en.tsv#L45-L53
[4] HIPE-data-v1.3-test-masked-bundle5-en.tsv#L56-L61
[5] https://en.wikipedia.org/wiki/Constitution_of_the_United_States
[6] HIPE-data-v1.3-test-en.tsv#L1663-L1665

2 Entity Linking for Historical Data

Entity linking (EL) is an information extraction task that semantically enriches documents by identifying pieces of text that refer to entities, and by matching each piece to an entry in a knowledge base (KB). Frequently, the detection of entities is delegated to an external named entity recognition (NER) system. Thus, in the state of the art, EL tools are either *end-to-end systems*, i.e. tools that perform both tasks, or *disambiguation systems* [11,18], i.e. tools that perform only the matching of entities and consider the first task as an input.

End-to-end EL systems were initially defined for contemporary documents [5]. First systems were focused on monolingual corpora and then gradually moved to a multilingual context. Some recent configuration, named Cross-Lingual Named Entity Linking (XEL), consist in analysing documents and named entities in a language different from the one used in the knowledge base. Some recent works proposed different XEL approaches: zero-shot transfer learning method by using a pivot language [27], hybrid approach using language-agnostic features that combine existing lookup-based and neural candidate generation methods [31], and the use of multilingual word embeddings to disambiguate mentions across languages [21].

Regarding the application of end-to-end EL in Digital Humanities, some works have focused on using available EL approaches to analyse historical data [16,23,28]. Other works have concentrated on developing features and rules for improving EL in a specific domain [13] or entity types [30,3,4]. Furthermore, some researchers have investigated the effect of issues frequently found in historical documents on the task of EL [13,20].

Some NER and EL systems dedicated to historical documents have also been explored [16,23,24,28]. For instance, van Hooland *et al.* [16] 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 [28] used DBpedia Spotlight tool to disambiguate named entities on Bentham's manuscripts. Finally, Munnelly and Lawless [24] investigated the accuracy and overall suitability of EL systems in 17^{th} century depositions obtained during the 1641 Irish Rebellion.

Most of the developed end-to-end EL systems are monolingual like the work of Mosallam *et al.* [22]. The authors developed a monolingual unsupervised method to recognise person names, locations, and organisations in digitised French journals of the National Library of France (*Bibliothèque nationale de France*) from the 19^{th} 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 at al.[17] also analysed the French journal *Le Monde*'s 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 corresponding entities (Wikipedia article). Then, they linked these Wikipedia

articles to YAGO [26] to recognise and disambiguate entities in the archive of *Le Monde*.

Monolingual disambiguation systems have also been studied by focusing on specific types of entities in historical documents, e.g. person and place names. Smith and Crane [30] 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 19$^{\text{th}}$ century America. In order to overcome with the heterogeneous data and the mix of temporal references (e.g. places that changed their name through time), they proposed a method based on honorifics, generic geographic labels, and linguistic environments to recognise entities, while they made use of gazetteers, biographical information, and general linguistic knowledge to disambiguate these entities. Another work [3,4] focused on authors' names in French literary criticism texts and scientific essays from the 19$^{\text{th}}$ and early 20$^{\text{th}}$ 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. Then, it crawls data from other linked data sets using equivalence links and fuses graphs of homologous individuals into a non-redundant graph in order to select the best candidate.

Heino *et al.* [13] investigated EL in a particular domain, the Second World War in Finland, using the reference datasets of WarSampo. 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.

The impact of OCR errors on EL systems, to our knowledge, has rarely been analysed 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.* [20], reported that EL systems for contemporary documents can see their performance decreased 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 knowledge base. Our approach contains several techniques to reduce the impact of the problems generated by the historical data issues, e.g. multilingualism, grammatical errors generated by OCR engines, and linguistic variation over time.

3 Historical Datasets

Unlike contemporary data that have multiple EL resources and tools, historical documents face the problem of lacking annotated resources. Moreover, contemporary resources are not suitable to build accurate tools over historical data due to the variations in orthographic and grammatical rules, not to mention the fact that names of persons, organisations, and places could have significantly changed over time.

To the best of our knowledge, there are few publicly available corpora in the literature with manually annotated entities on historical documents. Most EL corpora are composed of contemporary documents. Unfortunately, they do not contain the distinctive features found in historical documents. In this work, we focus on two corpora that contain historical documents in English, Finnish, French, German, and Swedish.

The first corpus was produced for the CLEF HIPE 2020 challenge[7] [8]. This corpus is composed of articles published between 1738 and 2019 in Swiss, Luxembourgish, and American newspapers. It was manually annotated by native speakers according to HIPE annotation guidelines [8].

Table 1. Number of entities for the training, development, and test sets in CLEF HIPE 2020 and NewsEye corpora.

Split	**CLEF HIPE 2020**			**NewsEye**			
	German	**English**	**French**	**German**	**Finnish**	**French**	**Swedish**
training	3,505	-	6,885	–	1,326	–	1,559
development	1,390	967	1,723	–	284	–	335
test	1,147	449	1,600	7,349	287	5,090	337

The second corpus was produced for the Horizon 2020 NewsEye project[8] and it is a collection of annotated historical newspapers in French, German, Finnish, and Swedish. These newspapers were collected by the national libraries of France[9] (BnF), with documents from 1814 to 1944, Austria[10] (ONB) with documents from 1845 to 1945, and Finland[11] (NLF), with Finnish and Swedish documents from 1771 to 1910 and 1920, respectively.

Both corpora contain named entities that are classified according to their type and, when possible, linked to their Wikidata ID. Non-existent entities in the Wikidata KB are linked to NIL entries. Table 1 shows the statistics of the datasets for the training, development, and test partitions.

4 Multilingual End-to-end Entity Linking

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

[7] https://impresso.github.io/CLEF-HIPE-2020/
[8] https://www.newseye.eu
[9] https://www.bnf.fr
[10] https://www.onb.ac.at
[11] https://www.kansalliskirjasto.fi

4.1 Building Resources

By definition of the task, EL systems use knowledge bases (KB) as entry reference but their use is not limited to it. KBs are also used by EL systems for tasks such as extraction of supplementary contexts or surface names, disambiguation of cases, or linking of entities with a particular website entry. In the following paragraphs, we present the most representative KBs used in this domain.

Wikipedia[12], a multilingual encyclopedia available in 285 languages, is commonly used as KB in the state-of-the-art. For instance, [11,18] make use of the English Wikipedia to disambiguate entity mentions in newspapers. Agirre *et al.*[1] used Wikipedia not only to disambiguate mentions found in historical documents but also to explore the feasibility of matching mentions with articles on Wikipedia according to their cultural heritage.

Wikidata[13] is a KB created by the Wikimedia Foundation[14] to store, in a structured way, data generated and used by the different Wikimedia projects, e.g. Wikipedia and Wiktionary. For instance, it has been used to annotate historical corpora, such as those used on this paper, CLEF HIPE 2020 and NewsEye.

DBpedia [19] is a KB that structures and categorise information collected from different Wikimedia projects, including Wikipedia and Wikidata, while including links to other KBs such as YAGO [26] or GeoNames[15]. For instance, it was used by [6] for annotating mentions of locations in *Historische Kranten*, a historical newspaper corpus. While [23] used DBpedia for annotating historical legal documents. Other examples of EL and DBpedia can be found in the works of [10,16].

In this work, we decided to build our own KB consisting of information from Wikipedia. Nevertheless, rather than just focusing on the English Wikipedia, we make use as well of the versions found in the languages used in the datasets to evaluate: French, German, Finnish, and Swedish. The reasoning behind this is that despite the richness and coverage of the English Wikipedia, on occasion other versions of Wikipedia might contain information that is only found in a specific language. For instance, *Valentin Simond*, owner of the French newspaper *L'Écho de Paris*, has an entry only in the French Wikipedia[16].

4.2 Entity Embeddings

Based on the work of [11], we decided to create entity embeddings for each language by generating two conditional probability distributions. The first one, the "positive distribution", is a probability approximation based on word-entity co-occurrence counts, i.e. which words appear in the context of an entity. The counts were obtained, in the first place, from the entity Wikipedia page, and,

[12] https://www.wikipedia.org
[13] https://www.wikidata.org
[14] https://www.wikimedia.org
[15] http://www.geonames.org
[16] https://fr.wikipedia.org/wiki/Valentin_Simond

in second place, from the context surrounding the entity in an annotated corpus using a fixed-length window. The second distribution, the "negative" one, was calculated by randomly sampling context windows that were unrelated to a specific entity. Both probability distributions were used to change the alignment of words embeddings with respect to an entity embedding. The positive probability distribution is expected to approach the embeddings of the co-occurring words with the embedding vector of the entity, while the negative probability distribution is used to distance the embeddings of words that are not related to an entity.

It should be noted that, unlike some works, where all the possible entities are known beforehand, in our work the creation of entity embeddings is not directed by a dataset. This is done to prevent bias and low generalisation. In case an entity does not have an entity embeddings, the EL system will propose a NIL.

4.3 Entity Disambiguation

The entity disambiguation model is based on the neural end-to-end entity linking architecture proposed by Kolitsas el al. [18]. The first advantage of this architecture is that it performs both entity linking and disambiguation. This method can then benefit from simplicity and from lack of error propagation. Furthermore, this architecture does not require complex feature engineering, which makes it easily adaptable to other languages.

For recognising all entity mentions in a document, Kolitsas *et al.* utilised an empirical probabilistic table entity−map, defined by $p(e|m)$. Where p is the probability of an entity e to be related to a mention m; $p(e|m)$ is calculated using the number of times that mention m refers e within Wikipedia. From this probabilistic table, it is possible to find which are the top entities that a mention span refers to.

The end-to-end EL model starts by encoding every token in the text input by concatenating word and character embeddings and fed into a Bidirectional Long Short Term Memory (BiLSTM) [14] network. This representation is used to project mentions of this document into a shared dimensional space with the same size as the entity embeddings. These embeddings are fixed continuous entity representations generated separately, namely in the same manner as presented in [11], and aforementioned in Subsection 4.2. In order to analyse long context dependencies of mentions, the authors utilised the attention mechanism proposed by [11]. This mechanism provides one context embedding per mention based on surrounding context words that are related to at least one of the candidate entities.

The final local score for each mention is determined by the combination of the $\log p(e|m)$, the similarity between the analysed mention and the candidate entity, and the long-range context attention for this mention. Finally, a top layer in the neural network promotes the coherence among disambiguated entities inside the same document.

4.4 Match Corrections

Multiple EL approaches, 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, it cannot propose candidates. In historic documents, not matching entities is a frequent problem, due to their inherent nature and processing, as explained in Section 1.

To increase the matching of entities in the probability table, we propose an analysis that consists of exploring several surface name variations using multiple heuristics. For instance, we evaluate variations by lower and uppercasing, capitalising words, concatenating surrounding words, removing stopwords, and transliterating special characters, like accentuated letters, to Latin characters. If after applying the previous heuristics, a match is still lacking, we use the Levenshtein distance to overcome more complex cases, such as spelling mistakes or transcription errors generated by the OCR systems.

4.5 Multilingualism

Historical and literary documents may contain words and phrases in a language different from that of the document under analysis. For instance, as shown in Figure 1(d), an English article uses "Porte de Namur" instead of "Namur Gate". However, the former only exists in the French probability table while the latter is only found in the English one. To overcome this problem, we combined the probability tables of several languages in order to identify the surface names of entities in multiple languages.

4.6 Filtering

To improve the accuracy of the candidates provided by the EL systems, we use a post-processing filter based on heuristics and DBpedia. Specifically, we utilise DBpedia's SPARQL Endpoint Query Service[17]. This filter uses DBpedia's hierarchical structure for specifying categories that represent each named entity type. For instance, entities belonging to a location type were associated with categories such as "dbo:Location" and "dbo:Settlement". The categories associated with each entity type were manually defined. Specifically, after requesting to the EL system the top five candidates for each named entity, the filtering steps are the following:

1. Verify that each candidate is in DBpedia and is associated with the correct categories. Candidates not matching the categories are put at the bottom of the rankings after a NIL;
2. Request to DBpedia the name of the candidates in the language of analysis; if the named entity is of type person, request as well the year of birth;
3. (Only if available) Remove those candidates that were born 10 years after the document publication;

[17] https://wiki.dbpedia.org/public-sparql-endpoint

4. Among the candidates with a retrieved name, find the most similar with respect to the named entity using Fuzzy Wuzzy Weighted Ratio[18];
5. The most similar candidate is ranked at the top;
6. If the ranking does not contain a NIL, add one as the last possible candidate.

Since DBpedia does not always contain the requested candidate or the candidate's name, we rely as well on DBpedia Chapters when available. For instance, "Turku" is categorised in DBpedia[19] but its name in Swedish, "Åbo" is not indexed; nevertheless, its Swedish name can be found in the Swedish DBpedia Chapter[20]. Another example is the case of "Luther-Werke", which does not exist in DBpedia, but it does exist in the German DBpedia Chapter[21].

5 Experimental Settings

In the context of multilingual historical newspapers, documents tend to contain local information that is often specific to a language and one or more related geographical areas. The use of KB in the historical newspaper's language is an obvious choice because it reduces problems of data consistency while decreases noise from entities in other languages. For instance, entities can represent different things according to each KB. For example, the English and the Finnish Wikipedia pages with the title "Paris" do not describe the same entity; in Finnish "Paris" make reference to Greek mythology while the French capital is known as "Pariisi". Therefore, we trained our EL model for the corresponding language of historical newspapers.

For the entity embeddings and the entity disambiguation model, we used the pre-trained multilingual MUSE[22] word embeddings with of size 300 for all the languages in the corpora. The character embeddings are of size 50. As no historical data is available for English, we used the AIDA dataset [15] and validated on the CLEF HIPE 2020 data. Based on the statistical analysis of the training data, we defined a Levenshtein distance ratio of 0.93 to search for other mentions in the probability table if this mention does not have a corresponding entry in the table[23].

For the evaluation, we compute precision (P), recall (R), and F-score (F1) measures calculated on the full corpus (micro-averaging). For the mentions without corresponding entries in the KB, EL systems provide a NIL entry to indicate that these mentions do not have a ground-truth entity in the KB.

[18] https://github.com/seatgeek/fuzzywuzzy
[19] http://dbpedia.org/page/Turku
[20] http://sv.dbpedia.org/page/%C3%85bo
[21] http://de.dbpedia.org/page/Luther-Werke
[22] https://github.com/facebookresearch/MUSE
[23] The source code of our EL system is available at: https://github.com/NewsEye/Named-Entity-Linking/tree/master/multilingual_entity_linking

6 Evaluation

As we previously stated, the semantic textual enrichment of historical documents depends on aspects such as the OCR quality or how a language has evolved. In order to analyse the EL performance on historical data and the impact of our techniques on the disambiguation of entities in historical data, we present in the Tables 2 and 3 a simple EL baseline ($p(e|m)$) and different combinations of our EL approach (henceforth MEL). For the filtering experiments (see Section 4.6), we predicted the five best candidate entities for a mention m based on the probability table ($p(e|m)$).

The configuration MEL+ML+MC+F[24] achieved the best results for French and German languages in CLEF HIPE 2020 corpora (Table 2).[25] Our model for English was trained on a contemporary dataset which degraded the performance of the MEL model and, consequently, all the variations. Despite the lack of historical training data, our model MEL+MC+F achieved the best results for the English data set (Table 2).

Table 2. Entity linking evaluation on the test CLEF HIPE 2020 data

Methods	English			French			German		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
$p(e\|m)$	0.595	0.593	0.594	0.586	0.583	0.585	0.532	0.530	0.531
MEL	0.549	0.546	0.547	0.535	0.532	0.533	0.484	0.482	0.483
MEL+F	0.608	0.607	0.607	0.591	0.588	0.59	0.528	0.528	0.528
MEL+ML	0.535	0.533	0.534	0.554	0.551	0.552	0.492	0.49	0.491
MEL+ML+F	0.595	0.593	0.594	0.602	0.600	0.601	0.538	0.537	0.538
MEL+MC	0.559	0.557	0.558	0.556	0.553	0.555	0.500	0.498	0.499
MEL+MC+F	**0.613**	**0.613**	**0.613**	0.621	0.619	0.620	0.538	0.537	0.538
MEL+ML+MC	0.547	0.546	0.547	0.577	0.574	0.576	0.507	0.505	0.506
MEL+ML+MC+F	0.589	0.589	0.589	**0.630**	**0.628**	**0.629**	**0.557**	**0.556**	**0.557**

ML: Multilingualism; MC: Match correction; F: Filter

For the NewsEye corpora, the MEL+MC+F version achieved the best results for all languages (Table 3). Similar to CLEF HIPE 2020, the MEL version generated the worst predictions. The filter increased the F-scores values of all EL versions. The combination of probability tables had almost no changes in the predictions.

[24] The MEL+ML+MC+F model (team 10-run 1) [2] achieved the best performance for almost all metrics in English, French, and German on the CLEF HIPE 2020 shared task results.

[25] The filter used in CLEF HIPE 2020 was modified in this work to improve accuracy and support DBpedia Chapters.

Table 3. Entity linking evaluation on the test NewsEye data

Methods	Finnish			French			German			Swedish		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
$p(e\|m)$	0.522	0.500	0.511	0.579	0.587	0.583	0.596	0.601	0.599	0.473	0.479	0.476
MEL	0.495	0.471	0.483	0.554	0.556	0.555	0.579	0.575	0.577	0.388	0.392	0.39
MEL+F	0.515	0.490	0.502	0.588	0.601	0.594	0.588	0.601	0.594	0.487	0.494	0.491
MEL+ML	0.505	0.481	0.493	0.555	0.558	0.557	0.575	0.573	0.574	0.392	0.397	0.394
MEL+ML+F	0.486	0.471	0.479	0.586	0.601	0.593	0.586	0.601	0.593	0.491	0.499	0.495
MEL+MC	0.501	0.481	0.491	0.562	0.568	0.565	0.582	0.580	0.581	0.386	0.390	0.388
MEL+MC+F	**0.527**	**0.502**	**0.515**	**0.597**	**0.611**	**0.604**	**0.597**	**0.611**	**0.604**	**0.513**	**0.521**	**0.517**
MEL+ML+MC	0.504	0.486	0.495	0.564	0.570	0.567	0.578	0.577	0.577	0.386	0.392	0.389
MEL+ML+MC+F	0.500	0.481	0.490	0.595	0.611	0.602	0.595	0.611	0.602	0.511	0.519	0.515

ML: Multilingualism; MC: Match correction; F: Filter

Though we generated the embedding representation for the 1.5M most frequent entities in each Wikipedia language, several historical entities are not so frequent on this KB. As our EL approach only disambiguates candidate entities that contain embedding representations, the MEL version achieved worse results than the baseline ($p(e|m)$). The major impact of this limitation was on the CLEF HIPE 2020 corpora where our approach had a drop of 0.05 in the F-score values.

Multilingualism The combination of probability tables of several languages has slightly improved the results on both corpora. This combination provided different surface names for an entity in different languages. In addition, this combination of probability tables allowed our models to disambiguate entities that are non-existent in some KBs. For example, the Russian politician "Nikolaï Alexeïevitch Maklakov" who is mentioned in the Finnish data does not exist in our Finnish KB, but he exists in our English and French KBs.

Despite providing additional surface variations, some surface names (e.g. acronyms) can have different meanings in different languages. Other potential risks are mentions with some OCR mistakes that can make reference to another entity in other languages and the combination of probability tables can increase the number of candidate entities and the ambiguity of mentions.

Match Corrections Our different analysis to normalise mentions and correct small mistakes generated by the OCR engine improved the performance of our approach. CLEF HIPE 2020 benefited sightly more from this technique than NewsEye. This could be either due to differences in the images quality, type of OCR used or manual correction.

On one hand, the combination of normalisation and Levenshtein distance methods allowed our method to correct mentions like "Londires" and "Toujquet" to "Londres" and "Touquet", respectively. On the other hand, our method could not find the correct mentions for simple cases. In the example "Gazstte of the Unites States", our approach did not find corresponding candidates for this mention. The correct answer is "Gazette of the United States"; however, the Levenshtein distance ratio is 0.928 and our threshold to correct a mention is

0.93. Another example of OCR errors is the mention "United Stares". In this case, the correct entity is "United States"; however, the candidate mention in the probability with the best Levenshtein distance ratio is "United Stars" which made our approach generated the wrong disambiguation. A lower Levenshtein distance ratio may find more degraded mention; however, this low ratio can generate too many mistakes for entities that not exist in KB. In the future, we will explore whether Fuzzy Wuzzy, an improved Levenshtein distance used in the filter (Section 4.6), could alleviate these issues.

Filtering The use of a post-processing filter for refining the top five most probable candidates, allowed us to achieve the best results, as observed in Table 2 and Table 3. Specifically, with the filter, we prioritised the candidates that not only were the most similar to the named entity but also, those that agreed with the named entity type and publication year. For instance, in an English newspaper published in 1810 the named entity of type person "Mr. Vance"[26] had for candidates the following Wikidata IDS: "Q507981" (location), "Q19118257" (person born in 1885), "Q985481" (location), and "Q7914040" (person born in 1930). Thanks to the filter, we observed that most of the candidates belonged to locations, while the proposed people were born long after the journal publication; thus, the best candidate should be a NIL, which in fact was the correct prediction. Despite DBpedia does not support languages such as Finnish, the filter can still improve the results using only the information regarding named entity categories, as seen in Table 3. It should be noticed that the filter is not free of errors. In some cases, the best candidate was positioned at the end of the rankings because DBpedia's categories did not match the categories defined for the named entity type, e.g. the journal "Le Temps", a product-type named entity, is not classified as a human work in DBpedia[27].

As digital library frameworks tend to provide the top N most probable entities for a mention in a context, we analysed the performance of the best two EL approach versions when we provide the top three candidate entities for each mention. These results are presented in Table 4. The MEL+MC+F method achieved the best average F-score, which is remarkable considering that the issues encountered in multilingual historical data can increase the difficulty of this task. Compared to Tables 2 and 3, the results are at least 14% better than the top one prediction.

Table 4. F-scores values for the top three candidate entities on the test data sets.

Methods	**CLEF HIPE 2020**			**NewsEye**			
	English	**French**	**German**	**Finnish**	**French**	**German**	**Swedish**
MEL+MC+F	**0.726**	**0.691**	0.623	**0.598**	0.706	0.699	0.594
MEL+ML+MC+F	0.710	0.690	**0.645**	0.566	**0.710**	**0.700**	**0.605**

ML: Multilingualism; MC: Match correction; F: Filter

[26] HIPE-data-v1.3-test-en.tsv#L4232-L4234

[27] http://dbpedia.org/page/Le_Temps_(Paris)

Based on all the previous results, we can observe that our EL approach outperformed the baseline for both corpora in all languages. Thus, we can conclude that the proposed techniques partially attenuated the impact of historical data issues. As well, the proposition of the best candidates can accelerate the work of librarians and humanities professionals in the analysis of historical documents in several languages and on different subjects. Finally, despite the recent progress, the EL for historical data is still a challenging task due to the multiple constraints. Examples of these limitations are the lack of annotated training data and the existence of multiple missing historical entities in the KBs, which can limit the training of more robust models.

7 Conclusion

Historical documents are essential resources for cultural and historical heritage. Enriching semantically historical documents, with aspects such as named entity recognition and entity linking, can improve their analysis and exploitation within digital libraries. In this work, we investigated a multilingual end-to-end entity linking system created for processing historical documents and disambiguate entities in English, Finnish, French, German, and Swedish. Specifically, we make use of entities embeddings, built from Wikipedia in multiple languages, along with a neural attention mechanism that analyses context words and candidate entities embeddings to disambiguate mentions in historical documents.

Additionally, we proposed several techniques to minimise the impact of issues frequently found in historical data, such as multilingualism and errors related to OCR systems. As well, we presented a filtering process to improve the linking of entities. Our evaluation on two historical corpora (CLEF HIPE 2020 and NewsEye) showed that our methods outperform the baseline and considerably reduce the impact of historical document issues on different subjects and languages.

There are several potential avenues of research and application. Following the idea proposed by [7], entity linking in historical documents could be used to improve the coverage and relevance of historical entities within knowledge bases. Another perspective would be to adapt our entity linking approach to automatically generate ontologies for historical data. As well, it would be interesting to use diachronic embeddings to deal with named entities that have changed of name through the time, such as "Beijing" in English[28]. Finally, we would like to improve our post-processing filter by including information from knowledge bases such as Wikidata or BabelNet [25].

Acknowledgments

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

[28] Google N-grams in English for "Beijing", "Peking", and "Pekin" between 1700 and 2008: books.google.com/ngrams/

References

1. 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)
2. 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: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Working Notes. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. CEUR-WS (2020)
3. Brando, C., Frontini, F., Ganascia, J.G.: Disambiguation of Named Entities in Cultural Heritage Texts Using Linked Data Sets. In: Morzy, T., Valduriez, P., Bellatreche, L. (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 (Sep 2015). https://doi.org/10.1007/978-3-319-23201-0_51, https://hal.archives-ouvertes.fr/hal-01203784
4. Brando, C., Frontini, F., Ganascia, J.G.: REDEN: Named Entity Linking in Digital Literary Editions Using Linked Data Sets. Complex Systems Informatics and Modeling Quarterly (7), 60 – 80 (Jul 2016). https://doi.org/10.7250/csimq.2016-7.04, https://hal.sorbonne-universite.fr/hal-01396037
5. 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 (Jun 2007), https://www.aclweb.org/anthology/D07-1074
6. 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 International Publishing (2015). https://doi.org/10.1007/978-3-319-23201-0_50
7. Dredze, M., McNamee, P., Rao, D., Gerber, A., Finin, T.: Entity disambiguation for knowledge base population. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). pp. 277–285. Coling 2010 Organizing Committee, Beijing, China (Aug 2010), https://www.aclweb.org/anthology/C10-1032
8. Ehrmann, Romanello, Clematide, Flückiger: HIPE - Shared Task Participation Guidelines (Jan 2020). https://doi.org/10.5281/zenodo.3677171, https://doi.org/10.5281/zenodo.3677171
9. Ehrmann, M., Romanello, M., Bircher, S., Clematide, S.: Introducing the CLEF 2020 HIPE shared task: Named entity recognition and linking on historical newspapers. In: Jose, J.M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M.J., Martins, F. (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
10. 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 (06 2015)
11. 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

12. Gefen, A.: Les enjeux épistémologiques des humanités numériques. Socio (2015). https://doi.org/https://doi.org/10.4000/socio.1296
13. 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 International Publishing, Galway, Ireland (2017). https://doi.org/10.1007/978-3-319-59888-8_10
14. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
15. 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. (Jul 2011), https://www.aclweb.org/anthology/D11-1072
16. 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 in the Humanities **30**(2), 262–279 (11 2013). https://doi.org/10.1093/llc/fqt067
17. Huet, T., Biega, J., Suchanek, F.M.: Mining history with le monde. In: Proceedings of the 2013 Workshop on Automated Knowledge Base Construction. p. 49–54. AKBC '13, Association for Computing Machinery, New York, NY, USA (2013). https://doi.org/10.1145/2509558.2509567
18. 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
19. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., Kleef, P.v., Auer, S., Bizer, C.: DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web Journal **6**(2), 167–195 (2015). https://doi.org/10.3233/SW-140134
20. 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
21. 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. p. 329–332. JCDL '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3383583.3398597, https://doi.org/10.1145/3383583.3398597
22. 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 International Publishing, St. Petersburg, Russia (2014). https://doi.org/10.1007/978-3-319-08976-8_2
23. 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. p. 59–68. JCDL'18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3197026.3197055

24. 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
25. Navigli, R., Ponzetto, S.P.: BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence **193**, 217–250 (2012). https://doi.org/10.1016/j.artint.2012.07.001
26. Pellissier Tanon, T., Weikum, G., Suchanek, F.: YAGO 4: A reason-able knowledge base. In: Harth, A., Kirrane, S., Ngonga Ngomo, A.C., Paulheim, H., Rula, A., Gentile, A.L., Haase, P., Cochez, M. (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
27. 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 (January 2019). https://doi.org/10.1609/aaai.v33i01.33016924
28. Ruiz, P., Poibeau, T.: Mapping the Bentham Corpus: Concept-based Navigation. Journal of Data Mining and Digital Humanities. **Special Issue: Digital Humanities between knowledge and know-how (Atelier Digit_Hum)** (Mar 2019), https://hal.archives-ouvertes.fr/hal-01915730
29. Shen, W., Wang, J., Han, J.: Entity linking with a knowledge base: Issues, techniques, and solutions. IEEE Transactions on Knowledge and Data Engineering **27**(2), 443–460 (2015). https://doi.org/10.1109/TKDE.2014.2327028
30. 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. p. 127–136. ECDL '01, Springer-Verlag, Darmstadt, Germany (2001). https://doi.org/10.1007/3-540-44796-2_12
31. 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 (Nov 2019). https://doi.org/10.18653/v1/D19-6127