

The Past is a Foreign Place: Improving Toponym Linking for Historical Newspapers

Mariona Coll Ardanuy^{1,2,*,†}, Federico Nanni^{1,*}, Kaspar Beelen^{1,3,†} and Luke Hare^{1,†}

¹The Alan Turing Institute, British Library, London, United Kingdom

²PRHLT Research Center, Universitat Politècnica de València, València, Spain

³Digital Humanities Research Hub, School of Advanced Study, Senate House, London, United Kingdom

Abstract

In this paper, we examine the application of toponym linking to digitised historical newspapers. These collections constitute the largest trove of historical text data available to researchers in the humanities. They contain varied, fine-grained information about the past, anchored in a specific place and time. Place names (or toponyms) are common entry points for starting exploring these collections. In this paper, we introduce a new tool for toponym linking and resolution, *T-Res*, a modular, flexible, and open-source pipeline, which is built on top of robust state-of-the-art approaches. We present a comprehensive step-by-step examination of this task in English, and conclude with a case study in which we show how toponym linking enables historical research in the digitised press.

Keywords

toponym resolution, entity linking, historical newspapers, nineteenth-century, toponym linking

1. Introduction

The digitised press is one of the largest historical collections available to humanities researchers; it is an invaluable source for better understanding past and present society. Initiatives to digitise newspaper collections have emerged all over the globe and continue to grow today [3]. Moreover, advances in data science and natural language processing have enabled researchers in the (computational) humanities to interrogate these massive data sets in previously unimaginable ways. Applying powerful computational tools to such massive data sets has the potential to fundamentally transform historical research.

Entity linking (EL) occupies a central position in large-scale text-based interdisciplinary research. Being able to automatically identify entities and link them to a predefined knowledge base (KB) opens up novel avenues for exploration and analysis. However, as research has regularly noted [42, 32, 12], EL systems still encounter many difficulties when processing historical

CHR 2023: Computational Humanities Research Conference, December 6 – 8, 2023, Paris, France


*Corresponding authors.


†Work conducted while at The Alan Turing Institute.

‡Contributions of each author: *Methodology*: MCA, FN, KB; *Implementation*: MCA, FN; *Reproducibility*: FN, LH, MCA; *Application and case studies*: KB; *Analysis and experiments*: MCA, FN; *Writing*: MCA, FN, KB.

✉ mcoll@prhl.upv.es (M. Coll Ardanuy); fnanni@turing.ac.uk (F. Nanni); kaspar.beelen@sas.ac.uk (K. Beelen); lhare@turing.ac.uk (L. Hare)

ORCID 0000-0001-8455-7196 (M. Coll Ardanuy); 0000-0003-2484-4331 (F. Nanni); 0000-0001-7331-1174 (K. Beelen)

 © 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

newspaper texts. In this paper, we focus on the specific task of identifying and linking geographical named entities (i.e. toponyms) in historical newspapers in English. This task presents certain additional challenges to standard EL, as illustrated with the following news fragments:

- *CxitiTCHUßCit, June 10—Yesterday being the day appointed for the election of taro gentlemen to tepcoeot this borough in the new imperial Parliament [...].*¹
- *Leghorn, April 6. ETTERS from Condantinopte, dated March 3, mention, tliat an Earthquake had lately hapL 3 pened at Tauris, the Capita! of the Province of *IS)ra Ariherbigan, in Percfia.*²

Most visible are the errors introduced during digitisation and optical character recognition (OCR). Such errors can occur both in the named entity (sometimes even rendering it incomprehensible for the human reader) or in the context of the entity. Secondly, historical newspapers portray a world that has changed, while at the same time being often very regional in their focus [20]: in the first example, ‘*CxitiTCHUßCit*’ (i.e. Christchurch) refers to the town in Dorset—which would have been the first reference of the readers of *The Dorset County Chronicle*—instead of the (today) more well-known city in New Zealand. Despite their strong regional focus, most publications also covered international news, reflecting a state (and vision) of the world that has changed: notice the use of the toponyms ‘*Leghorn*’ for Livorno, ‘*Condantinopte*’ (i.e. Constantinople) for Istanbul, and ‘*Tauris*’ for Tabriz, capital of ‘**IS)ra Ariherbigan*’ (i.e. East Azerbaijan) in ‘*Percfia*’ (i.e. Persia, modern Iran).

In this paper, we perform a comprehensive step-by-step examination of toponym linking in the historical newspapers domain in English. As a result of this analysis, we present *T-Res*,³ a new tool for toponym linking and resolution of historical newspapers in English, built on top of existing robust technologies, such as transformers [49] for fine-tuning a BERT language model for named entity recognition [11]; *DeezyMatch* [23] for candidate selection; and the work of Le and Titov [27] and Ganea and Hofmann [17] for entity disambiguation, via the Radboud Entity Linker (REL) implementation [24]. *T-Res* has been developed to assist researchers explore large collections of digitised historical newspapers, and has been designed to tackle common problems of working with these data. It is implemented as a modular pipeline, and is both user-friendly and flexible, where the user can either provide their own resources and datasets and train their own models, or they can load existing models. We conclude our paper with a preliminary but realistic case study in which we showcase how *T-Res* can be used to support historical research.

2. Related Work

Entity linking (EL) is often treated as a three-step process: (1) named entity recognition (NER) is the task of detecting mentions, (2) candidate selection (CS) is the task of selecting a subset of potential referents from a knowledge base (KB) for the detected mentions, and (3) entity disambiguation (ED) finds the best match, if any, from the pool of selected candidates. EL benchmarks in English consist mainly of texts from the general domain, which mostly feature

¹ *The Dorset County Chronicle*, 1864-04-14.

² *The Manchester Mercury*, 1780-05-30.

³ <https://github.com/Living-with-machines/T-Res>.

prominent entities [48]. Therefore, tools that perform well on such datasets, are often found to deteriorate in other domains, such as on historical documents [38, 42, 36, 32]. The HIPE 2020 shared task⁴ [14] was created to address some of the EL challenges that are specific to digitised historical documents.

Historical digitised data has certain traits that are typically absent from standard EL benchmarks [13]. The presence of OCR errors is a persistent problem. In their assessment of the impact of OCR in downstream tasks, Strien, Beelen, Coll Ardanuy, Hosseini, McGillivray, and Colavizza [46] and Hamdi, Jean-Caurant, Sidère, Coustaty, and Doucet [19] observe how NER performance decreases as text quality declines. The results of the HIPE-2020 shared task [14] (and its continuation HIPE-2022 [15]) point to the importance of having in-domain training data for NER, suggesting that fine-tuning on noisy data results in better performance on similarly noisy data. Similarly, Manjavacas and Fonteyn [31] show how NER models perform better when they have been fine-tuned on top of base models which were originally pre-trained on in-domain data, in this case, historical digitised texts. González-Gallardo, Boros, Girdhar, Hamdi, Moreno, and Doucet [18] evaluated the performance of OpenAI’s ChatGPT on the task of detecting (in a zero-shot manner) named entities in historical documents, revealing that, similarly, ChatGPT struggles with identifying entities in OCR text.⁵

Candidate selection is the least studied of the three sub-tasks. The identification of potential candidates from the KB (usually based on collaboratively-built resources such as Wikipedia, Wikidata, of Freebase) has traditionally been approached by performing exact or partial string matching between a mention and the entries in the KB [33, 45]. Since most popular EL benchmarks consist of very clean text, this step does not often pose an obstacle for achieving a good EL performance. In other words, plain string matching goes a long way. However, when working with noisy text, basic string matching is far from sufficient [51]. In the domain of digitised historical newspapers, Linhares Pontes, Cabrera-Diego, Moreno, Boros, Hamdi, Doucet, Sidere, and Coustaty [29] propose a series of pre-processing heuristics used in combination with a post-correction step, based on mappings of common OCR errors observed in the data. Traditional fuzzy string matching techniques based on edit distance (such as Levenshtein) can deal quite accurately with OCR text, but they are not a viable solution for real-time EL, since they are computationally inefficient [10]. DeezyMatch [23], a software library for neural fuzzy string matching, was developed as a response to this problem, building on Santos, Murrieta-Flores, Calado, and Martins [43].

The last step of the pipeline is a disambiguation task, consisting of selecting the most appropriate entity from the pool of previously selected potential candidates. The entity disambiguation literature often distinguishes between local models—which rely only on the mention’s context and the entities’ priors, often based on hyperlink counts from large resources such as Wikipedia [8, 35, 34]—and global models—which take interdependencies between entities into account [41, 25, 27], with the more recent approaches learning deep representations for relations between entities and mentions. In the domain of historical newspapers, Boros, Pontes, Cabrera-Diego, Hamdi, Moreno, Sidère, and Doucet [6] and Linhares Pontes, Hamdi, Sidere, and Doucet [30] build on these approaches, and emphasise the importance of good knowledge

⁴<https://impresso.github.io/CLEF-HIPE-2020/>.

⁵Our own experiments with ChatGPT were not more successful.

representation.

EL pipelines encapsulate all steps in one toolkit. DBpedia Spotlight [33] and TagMe! [16, 40] are two of the first and most widely used out-of-the-box linkers. More recently, REL [24] was developed to overcome some of the shortcomings of previous systems, building on state-of-the-art approaches. REL uses Flair [1] for recognition. For candidate selection, just like most other state-of-the-art approaches, REL employs a series of string-based heuristics to find potential candidates, which are ranked according to a combination of entity priors and a measure of similarity between the entity and the context of the mention, as in Ganea and Hofmann [17], using Wikipedia2Vec’s [50] word and entity embeddings. The local coherence between mention and entity is computed as defined in Ganea and Hofmann [17] and uses the global disambiguation strategy proposed in Le and Titov [27]. REL is fast, user-friendly, easily customisable, and very well documented, including tutorials, examples and a running API.⁶

3. Terminology and Task Definition

In this paper, we use the terms *toponym linking* and (*geographic*) *entity linking* interchangeably to refer to the end-to-end task of detecting mentions of places in texts and linking them to their referent in a knowledge base.⁷ Formally defined, given a document D , the goal is to detect mentions of places m_1, \dots, m_n and resolve them to their corresponding entities e_1, \dots, e_n in a knowledge base (KB). This is achieved in three steps. The first step, called *toponym recognition* or (*geographic*) *named entity recognition*, consists of detecting mentions of places m_1, \dots, m_n in a document D . The second step is candidate selection, which, for a given mention m_i , aims at selecting a subset of potential k entities $C_i = (e_{i1}, \dots, e_{ik})$ from the KB. The last step is *entity disambiguation*, which, given the set of candidates C_i for mention m_i , consists of selecting the candidate that is the correct entity e_i for mention m_i , or return NIL if there is none. Finally, we define *toponym resolution* as the task of retrieving the coordinates of the predicted entities.

4. Experiments

4.1. Knowledge Base

As is commonly done in entity linking, we have used Wikimedia resources (in this case, Wikipedia in combination with Wikidata) as the starting point for our KB. For each Wikipedia page (hereafter ‘entity’), we extracted all ways of referring to it over the entire Wikipedia collection by means of the anchor texts of the hyperlinks pointing to the page (hereafter ‘mentions’). We then mapped Wikipedia entities to Wikidata, and kept only the subset of entities that are

⁶See: <https://github.com/informagi/REL>. While more recent approaches have now surpassed it on the leaderboard, it is still positioned near the top, according to <https://paperswithcode.com/task/entity-linking>.

⁷Detecting and resolving mentions of places to their real world referents is a research problem shared by two different tasks: (1) Entity Linking, the task of linking named entities to their corresponding entries in a knowledge base, and (2) Toponym Resolution, also called geoparsing, the task of resolving place names to their spatial footprint, often their geographic coordinates [28]. Because of these slightly different objectives, both tasks are rarely evaluated jointly. We treat this problem as an EL task, in part because linking to Wikidata (instead of directly providing coordinates) gives the user access to other linked information.

geolocated on Wikidata.⁸ The resulting subset consists of 929,855 geolocated entities. In addition, for each entity, we keep the absolute and normalised mention-to-entity frequencies of all its mentions. Mention-to-entity frequencies are normalised per entity: for example, the settlement named ‘London’ in Kiribati has an absolute mention-to-entity count of 13 and a normalised frequency of 0.81 (the mention ‘London’ refers to the location in Kiribati 13 times, but the probability of London in Kiribati being referred to as ‘London’ is 0.81).

4.2. Datasets

We performed experiments on two different digitised historical newspaper datasets in English:

- *TopRes19th* (henceforth ***lwm***): This dataset was created by the *Living with Machines* project [9].⁹ In its latest version (v2), this dataset consists of 455 news articles in which places were manually annotated and linked to Wikipedia (which we have mapped to Wikidata). The news articles in this dataset were selected from local or regional newspapers based in different locations in England (Manchester, Ashton-under-Lyne, Poole and Dorchester), published between 1780 and 1870. In the dataset, toponyms are classified as ‘BUILDING’, ‘STREET’, ‘LOC’, ‘ALIEN’, ‘FICTION’, and ‘OTHER’, but the last three were found to occur between zero and five times in the whole dataset, therefore resulting negligible for training purposes. The dataset is split into training and test sets (343 and 112 articles respectively). We used 20% of the training set for development.
- *Hipe2020* (henceforth ***hipe***): This dataset was created by the *Impresso* project with data from the *Chronicling America* project, and was released as part of the HIPE2020 evaluation campaign on named entity processing on historical newspapers [14].¹⁰ It consists of news articles in English, French, and German. The English collection, which is the one we use, consists of 125 articles from 14 different newspapers (based in 14 different locations in the United States) published between 1790 and 1960. The named entities are manually identified and linked, whenever possible, to their corresponding Wikidata entity. While the dataset has other entity types (such as ‘person’ or ‘organisation’), in our experiments we consider only entities of the type ‘location’. This dataset does not have a training set, it is instead split into a development and test set (80 and 46 articles respectively).

4.3. Approaches

4.3.1. Named entity recognition

We fine-tuned a BERT model for token classification, using the *lwm* training set. We used a historical BERT model as base, *bert_1760_1900* [22], trained on books in English published

⁸We used the 2021-10-20 English Wikipedia version and the 2022-07-28 Wikidata version. We relied on the WikiExtractor (<https://github.com/attardi/wikiextractor>) tool to extract the content of each page from the Wikipedia XML dump, used WikiMapper (<https://github.com/jcklie/wikimapper>) to map Wikipedia titles to Wikidata QIDs, and used the Wikidata property P625 to filter out non-geographic entities.

⁹The *lwm* dataset is available at <https://doi.org/10.23636/r7d4-kw08> (CC BY-NC-SA 4.0). Newspaper data was provided by Findmypast Limited from the British Newspaper Archive, a partnership between the British Library and Findmypast: <https://www.britishnewspaperarchive.co.uk/>.

¹⁰The *hipe* dataset is available at <https://zenodo.org/record/6046853> (CC BY-NC-SA 4.0).

between 1760 and 1900.¹¹ To fine-tune for toponym recognition, we used a learning rate of 0.00005, a batch size of 8, 10 epochs, and weight decay of 0.00.¹² We perform a series of post-processing steps to fix obvious mistagging errors: we corrected I- labels at the beginning of a new entity, removed inner I- tags due to nested entities, and fixed prefix assignment errors in hyphenated entities.

4.3.2. Candidate selection

Our tool provides two main different strategies for candidate selection:¹³ one is based on exact matching, where candidates are retrieved from the KB if they are identical to the query; and the other is based on fuzzy string matching, using a deep learning approach to fuzzy string matching, DeezyMatch [23, 10] in a new fashion, which we expand on in the following paragraphs.

DeezyMatch for candidate selection DeezyMatch learns string transformations from a large set of positive and negative example pairs (e.g. both ‘Zuiich’ and ‘7urich’ are positive examples of OCR variations of ‘Zurich’, whereas ‘Munich’ is not). A model trained from these examples is then used to embed both (1) the query and (2) all name variations in the KB into vector representations. Candidate string variations are retrieved from the KB and ranked according to the similarity between their embedding representations and the query embedding.

We propose a new approach for generating positive and negative example pairs when large volumes of noisy text are available. We observed an interesting difference between static word embeddings learnt from clean text and learnt from OCR text. In the first case, as explained by the distributional hypothesis, the top nearest neighbours of a query tend to be words that are semantically similar. However, when word embeddings are trained on OCR text, many of the top nearest neighbours are OCR variations of the query. We used this observation to build a dataset of positive and negative matches from word2vec embeddings learnt from digitised English newspapers, where:

- If the string similarity between the nearest neighbour and the target word is high (such as ‘*maciine*’ and ‘*machine*’) and the nearest neighbour is not an existing word in English, we consider it an OCR string variation of the target word (i.e. positive example);
- If the string similarity between the nearest neighbour and the target word is low (such as ‘*maciine*’ and ‘*device*’), we consider it is not a string variation, as it is probably a synonym or near-synonym (i.e. negative example).

¹¹The model is available at https://huggingface.co/Livingwithmachines/bert_1760_1900.

¹²We selected these values based on previous research that performed a hyperparameter search for the same task and a different base model [44]. See more information at <https://github.com/dbmdz/clef-hipe/blob/main/experiments/clef-hipe-2022/>. The resulting toponym recognition model is available at <https://huggingface.co/Livingwithmachines/toponym-19thC-en>.

¹³We also provide functions for performing partial matching based on string overlap and fuzzy string matching based on the Damerau-Levenshtein edit distance. However, both methods are highly time-consuming, and therefore unusable for real-time scenarios.

We used openly-available word embeddings trained on digitised newspaper text from four different decades (1800s, 1830s, 1860s, and 1890s).¹⁴ Table 1 shows examples of positive and negative string matches generated with this approach. We expanded the resulting string pairs dataset by appending similar variations of place names obtained from our KB mention-to-entity mapping.¹⁵ The resulting dataset consists of 1,085,514 string pairs.

Table 1

Examples of positive and negative string matches for the target word ‘would’ obtained by filtering nearest neighbours in static word embeddings.

would	might	False
would	must	False
would	likely	False
would	woull	True
would	wonld	True
would	woubl	True

Candidate ranking and selection Given a query, the candidate selection step retrieves one or many potential name variations from the KB. In the exact match approach, only one name variation is retrieved (i.e. the identical match) with a similarity score of 1.0. In the DeezyMatch approach, the user can choose the number of name variations to retrieve and set the maximum accepted distance between the embeddings of the query and the KB mentions.¹⁶ The similarity score for each of the retrieved name variations is obtained from reverse-normalising the distance score against the threshold. Each name variation is then expanded to multiple Wikidata entities (i.e. candidates), using the mention-to-entity mapping from our KB.¹⁷

4.3.3. Entity disambiguation

The last step consists of finding the most likely entity from the pool of selected candidates for a given query. We provide two dummy baselines: the first baseline (*mostpopular*) selects the candidate which is more likely to be referred by a certain query, using the mention-to-entity absolute counts described in section 4.1. The second baseline (*bydistance*) is based on distance from the place of publication, in which the closest candidate from the place of publication is naively selected as the correct entity. Finally, our tool adopts REL’s entity disambiguation

¹⁴The word embeddings are available at <https://doi.org/10.5281/zenodo.7181682> (CC BY 4.0) [39]. For example, the nearest neighbours of ‘machine’ in word embeddings trained from digitised newspaper articles published in the 1860s are: ‘machines’, ‘maehine’, ‘maciine’, ‘machina’, ‘maohine’, ‘achine’, ‘miachine’, and ‘maohine’. We used the vocabulary of the 50d GLoVe embeddings to discern whether a word exists in English. Further details can be found in our GitHub repository.

¹⁵For example, the Wikidata entry Q7268098 is referred to as both ‘Qoorlugud’ or ‘Qorilugud’: they would be added as positive variations of each other.

¹⁶We selected one name variation per query, with an L2-norm similarity threshold set at 50.

¹⁷For example, given the query ‘Wiltshire’, the exact match approach would retrieve the mention ‘Wiltshire’ from the KB with a similarity score of 1.0, which would be expanded to the following Wikidata entities: Q23183, Q55448990, and Q8023421, since all of them are referred to as “Wiltshire” at least once in Wikipedia anchor texts.

implementation¹⁸ (*rel*), which is based on Ganea and Hofmann [17] and Le and Titov [27], and uses a neural approach to combining local mention-to-entity compatibility and global entity-to-entity coherence. We provide our own set of candidates (selected either with the *exact* or *deezy* match approach), which we pre-rank by averaging the string matching score and the relative mention-to-entity score, and the normalised absolute mention-to-entity score. We additionally provide the following two alterations to the REL disambiguation approach:

- **Providing information about the place of publication:** (*+publ*) Since we are aware of the strong local emphasis of the historical press, we experiment with artificially providing information on the place of publication (which is often available from the newspaper’s metadata) to the disambiguation module: we do so by adding one additional already-disambiguated entity per sentence, both in training and in testing, corresponding to the place of publication, and adding the publication place name also as part of the context of the sentence.
- **Unlinking micro locations:** (*:nil*) Streets and buildings have rather different characteristics than other locations typically found in news articles: they are often highly ambiguous and often entirely dependent on the cues provided by context. At the same time, they have a very limited coverage in Wikipedia (where only the most noteworthy streets or buildings are usually included). In this variation of the original method, only the LOC entities are disambiguated, whereas mentions classified as BUILDING or STREET are linked to NIL.

5. Evaluation and Discussion

In this section, we report and discuss the results assessed using the HIPE-scorer.¹⁹

5.1. Toponym Recognition

Table 2 shows the performance of our BERT-based historically-tuned NER approach on the *lwm* and *hipe* datasets. As a comparison, we provide the results obtained from running the REL API²⁰ (*rel*) and from the CLEF-HIPE-2022 shared task, including results from two participating teams (*aauzh* and *l3i*) and the neural baseline provided by the organisers (*neurbsl*).²¹ We report micro precision, recall, and F1-score on two different settings:

¹⁸We used the REL version at commit 9ca253b. We use the Wikipedia2Vec [50] word and entity embeddings shared by the authors, mapping them to Wikidata entities instead of Wikipedia titles.

¹⁹The HIPE-scorer (<https://github.com/hipe-eval/HIPE-scorer>, v1.1) is a Python module developed as part of the CLEF-HIPE-2020 evaluation campaign on named entity recognition and linking on historical newspapers.

²⁰REL’s default approach to recognising named entities in text uses Flair’s character-level sequence tagger [1], which is trained and evaluated on CoNLL-2003 data. Since REL tags not only locations, but also persons and organisations, we keep only those entities which are tagged as ‘LOC’ or whose prediction is an entity in our KB. REL returns a Wikipedia title, which we turn into a Wikidata QID.

²¹To learn more about the neural baseline and participating teams and approaches, read Ehrmann, Romanello, Najem-Meyer, Doucet, and Clemenide [15]. We only provide results for the ‘LOC’ tag for *hipe* because this dataset includes non-geographic entities. It is worth noting that there may be slight differences between the datasets used by us and those used in the shared task, because they have undergone different preparation steps. These differences are probably too small to be significant.

- *Strict*: exact boundary match, same entity type.
- *Type*: at least one token overlap, same entity type.

Note the considerable difference between the *strict* and *type* settings in all cases,²² the latter reflecting the correct identification of a mention’s presence while not agreeing on the exact named entity boundary. While the poorer performance of the out-of-the-box REL tool is not per se surprising (given that it has not been optimised for digitised historical text) the difference is substantial nonetheless. This experiment in fact highlights how, already at the recognition stage, there is a difference of around 15%–23% in terms of exact F1 between the out-of-the-box state-of-the-art method and a tool carefully attuned to the specific application domain. This is significant, since errors introduced in this step will percolate through the rest of the pipeline.

Table 2

Toponym recognition results for the *lwm* and the *hipe* datasets.

dataset	label	approach	<i>strict</i>			<i>type</i>		
			P	R	F1	P	R	F1
lwm	all	T-Res	0.821	0.834	0.828	0.856	0.870	0.863
		aauzh	0.816	0.760	0.787	0.869	0.810	0.838
		neurbsl	0.747	0.782	0.764	0.798	0.836	0.816
	loc	T-Res	0.856	0.871	0.863	0.882	0.897	0.890
		rel	0.577	0.707	0.636	0.601	0.737	0.662
	street	T-Res	0.815	0.815	0.815	0.870	0.870	0.870
	building	T-Res	0.638	0.649	0.644	0.718	0.730	0.724
hipe	loc	T-Res	0.658	0.676	0.667	0.775	0.797	0.786
		aauzh	0.683	0.547	0.607	0.807	0.646	0.718
		neurbsl	0.583	0.735	0.65	0.689	0.867	0.768
		l3i	0.714	0.691	0.702	0.806	0.779	0.792
		rel	0.449	0.599	0.513	0.556	0.742	0.635

5.2. Candidate Selection

In table 3, we report the highest possible performance that can be achieved during linking (i.e. *skyline*) by the different candidate selection strategies. In other words, a *skyline* true positive is when the correct entity has been selected as a potential candidate for a particular mention. The skyline, therefore, can be considered as a proxy for the quality of the different candidate selection approaches. We provide two evaluation settings: end-to-end EL (where mentions are identified using the best performing NER approach), and EL-only (where gold standard mentions are provided). In both cases, we report micro-scores using the ‘type’ evaluation setting, as finding the exact boundaries of the named entity is now not the goal. The results show the advantages of having a fuzzy string matching method (i.e. *deezy*), which in this case is trained on corpus-specific OCR variations, similar to those present in both datasets. The lower performance on *hipe* in the end-to-end EL setting is mostly just the consequence of the a worse toponym recognition in the previous step.

²²The smaller difference for our tool’s performance on *lwm* is expected since it uses the *lwm* training set for NER.

Table 3

Entity linking skyline results for *lwm* and *hipe* on (1) mentions detected using the best performing NER approach per dataset (*end-to-end EL*) and (2) the gold standard mentions (*EL-only*). We report micro precision, recall, and f1-score in the ‘type’ evaluation setting.

dataset	approach	<i>End-to-end EL</i>			<i>EL-only</i>		
		P	R	F1	P	R	F1
lwm	T-Res:exact	0.657	0.668	0.663	0.772	0.772	0.772
	T-Res:deezy	0.777	0.790	0.784	0.906	0.906	0.906
hipe	T-Res:exact	0.457	0.500	0.478	0.709	0.709	0.709
	T-Res:deezy	0.578	0.632	0.604	0.863	0.863	0.863

5.3. Entity Disambiguation

We report the results for the entity disambiguation step in Tables 4 and 5. The scores highlight how a very simple baseline—the combination of perfect match (at the selection stage) and most popular (at the disambiguation stage)—achieves a higher performance than the out-of-the-box REL system in the *lwm* dataset (but not on *hipe*), emphasising again the importance of a domain-specific module at the recognition stage.²³ The distance-based baseline, on the other hand, performs very poorly on both datasets. On the *lwm* dataset, the REL disambiguation approach (used as part of our tool), beats the *mostpopular* baseline, but not so in the *hipe* dataset, showing that the most common sense continues to be a very strong baseline. In both cases, interestingly, forcing streets and buildings to be ‘NIL’ results in a substantially higher performance. This suggests that most of these entities must be of type ‘NIL’ in the data (i.e. either too ambiguous to annotate or not present in the KB), but also that the disambiguation approach may not be suitable for these entities. While adding the place of publication has a positive impact on the *hipe* dataset, the impact on the *lwm* dataset is less clear.

Finally, we inspected more closely how the performance of our approaches vary based on several characteristics of our data. We split the *lwm* dataset into ten different subsets, each a unique combination of the decade in which the texts were written and the place of publication of the newspaper. We then performed a 10-fold validation of our results, where, in each fold, one subset was used for testing, another one for development, and the remaining eight subsets were used for training.²⁴ Detailed results are shown in Table 6.

First of all, we see a correlation between worse OCR quality (corresponding to the earlier data splits) and a lower skyline and linking performances. Second, there seems to be a correlation

²³Note that there are other factors that should also be taken into account. REL returns Wikipedia titles, which we mapped to Wikidata IDs, keeping only those results that are tagged as ‘LOC’ or can be mapped to geographic coordinates. However, it should be noted that REL uses its own KB, consisting not only of locations, therefore making the disambiguation a more difficult task since it is not only geographical entities that compete in the disambiguation process, but entities of any kind. It may be worth investigating, as part of future research, the impact this has on end-to-end EL. In providing this comparison, our goal is to illustrate the impact of using a general purpose EL system for this task, and stress the importance of developing tools that are targeted to the specific task and domain.

²⁴The ten subsets by publication are: Ashton-under-Lyne 1860, Dorchester 1820, Dorchester 1830, Dorchester 1860, Manchester 1780, Manchester 1800, Manchester 1820, Manchester 1830, Manchester 1860 and Poole 1860.

Table 4
Entity linking results for the *lwm* dataset.

Selection	Disambiguation	End-to-end EL			EL-only		
		P	R	F1	P	R	F1
rel-api	rel-api	0.459	0.498	0.478	—	—	—
T-Res:exact	mostpopular	0.552	0.561	0.557	0.637	0.637	0.637
	bydistance	0.170	0.172	0.171	0.215	0.215	0.215
T-Res:deezy	mostpopular	0.588	0.597	0.592	0.650	0.650	0.650
	bydistance	0.177	0.180	0.179	0.217	0.217	0.217
	rel	0.591	0.601	0.596	0.657	0.657	0.657
	rel:nil	0.652	0.663	0.658	0.741	0.741	0.741
	rel+publ	0.579	0.588	0.583	0.645	0.645	0.645
	rel+publ:nil	0.659	0.670	0.664	0.745	0.745	0.745

Table 5
Entity linking results for the *hipe* dataset.

Selection	Disambiguation	End-to-end EL			EL-only		
		P	R	F1	P	R	F1
rel-api	rel-api	0.365	0.489	0.418	—	—	—
T-Res:exact	mostpopular	0.377	0.412	0.394	0.588	0.588	0.588
	bydistance	0.085	0.093	0.089	0.220	0.220	0.220
T-Res:deezy	mostpopular	0.462	0.505	0.483	0.626	0.626	0.626
	bydistance	0.126	0.137	0.131	0.258	0.258	0.258
	rel	0.442	0.484	0.462	0.604	0.604	0.604
	rel:nil	0.447	0.489	0.467	0.599	0.599	0.599
	rel+publ	0.452	0.495	0.472	0.615	0.615	0.615
	rel+publ:nil	0.462	0.505	0.483	0.621	0.621	0.621

between the proportion of NILs and a lower median distance from publication, which is not entirely justified by a high presence of micro locations (i.e. streets and buildings), suggesting either (1) a higher difficulty for human annotators of finding the true referents of local mentions, or (2) the absence of local entities in the knowledge base. It is therefore not surprising to see *rel:nil* significantly improving on *mostpopular* in these cases, since the first maps buildings and streets to NIL. However, a closer inspection of the results also reveals the importance of the sensitivity of *rel:nil* (and *rel+publ:nil*) to context: for example, while “Ashton” is consistently resolved to be in Maryland by the *mostpopular* approach, it is in all but one case resolved adequately to Ashton-under-Lyne by the REL-based approaches.

5.4. Discussion and Limitations

Our research has focused on the geographic aspect of entity linking. However, T-Res could directly be used for general entity linking as well, with the exception of the *bydistance* linking

Table 6

Characteristics and entity linking results of each place-decade pair in the *lwm* dataset. We report the EL-only skyline and results for the *mostpopular* and *rel:nil*, using the *deezy* candidate selection approach. We also provide: (1) *OCR*, the mean of the per-word OCR confidence scores reported in the source metadata, (2) *Dist*, the median distance of the toponyms in relation to the publication place, (3) *NILs*, the proportion of cases in which the mention does not have a corresponding entity in the KB, and (4) *Micro*, the proportion of micro-toponyms (streets and buildings) in the split.

Publication	Decade	Characteristics of dataset				Approaches		
		OCR	Dist	NILs	Micro	skyline	mostpopular	rel:nil
Ashton	1860	0.892	9	0.32	0.298	0.958	0.551	0.824
Dorchester	1820	0.862	132	0.165	0.207	0.907	0.706	0.825
	1830	0.878	142	0.12	0.152	0.886	0.722	0.785
Manchester	1860	0.893	78	0.167	0.162	0.950	0.692	0.809
	1780	0.753	260	0.147	0.113	0.700	0.484	0.547
	1800	0.771	84	0.2	0.132	0.820	0.554	0.672
	1820	0.869	1134	0.113	0.124	0.880	0.707	0.770
	1830	0.884	211	0.193	0.311	0.951	0.672	0.792
Poole	1860	0.842	273	0.139	0.147	0.914	0.675	0.757
	1860	0.900	30	0.254	0.176	0.878	0.567	0.753

method, given a suitable knowledge base. We have focused on digitised historical newspapers, but T-Res could in principle be generalisable to other domains, possibly depending on additional annotated in-domain data. Further research is needed in this direction.

Our tool can be improved in many directions: each of its modules (named entity recognition, candidate selection, entity disambiguation) is open to improvements: for example, we can include more sophisticated NER fine-tuning strategies [7]. Computationally, however, the NER step is the clear bottle-neck of our tool: for example, it took about 90 minutes to recognise all toponyms in a sample of about 1,500 articles (4.2M of plain text) on a CPU, while the candidate selection and entity disambiguation steps (using *deezy* and *rel:nil+publ*) jointly took about three minutes.

By looking at our results from a more qualitative perspective, we realise that many of our errors stem from the KB itself. This is not surprising: not only does our KB mainly contain modern entities, it also uses modern relations between entities (via word and entity embeddings) as a way to represent their historical similarity. In the same vein as recent research [26, 5], we believe further research should go into understanding the impact of using domain appropriate entity embeddings in the disambiguation step, for example by training embeddings which take into account time and space.

Finally, another source of errors is the use of *DeezyMatch* for fuzzy string matching: while it allows us to efficiently discover entities which would otherwise have remained hidden, such as ‘Ashtonnder-Lyne’ for ‘Ashton-under-Lyne’ or ‘Horbury Junotion’ for ‘Horbury Junction’, its precision is lower than that of a traditional edit distance approach [10], sometimes resulting in what is called *hallucinations* in today’s AI jargon. For example, ‘Vieillevigine’ is matched to ‘Vielle Montaigne’. We therefore suggest combining the fast discoverability power of *Deezy*-

Match with a more conservative edit distance approach to filter obvious mismatches.²⁵ Finally, linking of micro locations is another direction that clearly requires further research.

6. Historical Case Study: Geographies from Below?

In this section, we present a case study as a type of user-testing and to assess how T-Res supports novel historical research on the digitised press. We explore the shifting geographies in Victorian working-class newspapers, analysing their local, national and transnational dimensions. For that, we have used the openly available British newspapers digitised by the *Living with Machines* project [47].²⁶ Motivated by a need to counterbalance the dominance of the liberal and conservative press with non-elite perspectives “from below” [4], the project prioritised the digitisation of “plebeian” newspapers channelling working-class voices, and selected exclusively ‘provincial’ papers to help research move beyond the traditional metropolitan emphasis in periodical research [21]. The corpus is not representative of the press (or society) as a whole, but it provides a solid starting point to explore the geographies embedded in the popular, working-class papers.

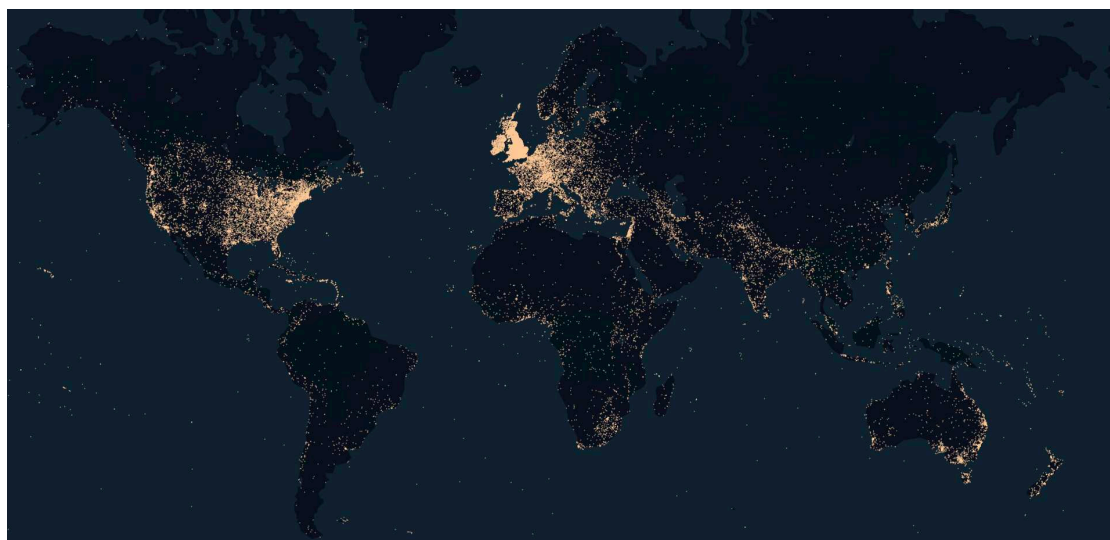


Figure 1: All unique places detected by T-Res (figure created with Kepler: <https://kepler.gl/demo>).

Our case study is based on a sample of more than 2,500 complete issues published between 1880 and 1900. This resulted in a set of 2.7 million detected toponyms. In the experiments, we kept only those georeferenced places classified as location (‘LOC’), which resulted in a collection of 1,770,412 data points comprising 46,820 distinct georeferenced locations. Figure 1 shows

²⁵In our case studies, we have applied a threshold of 0.85 edit distance similarity ratio between the mention and the returned DeezyMatch candidate, using the TheFuzz library: <https://github.com/seatgeek/thefuzz>.

²⁶<https://livingwithmachines.ac.uk/over-half-of-a-million-pages-of-historical-newspapers-now-openly-available/>

the global distribution of these unique toponyms. Below, we explore the places mentioned in the news across three different levels: the local, the national and the transnational. Firstly, we investigate to what extent the coverage of these late-Victorian newspapers was limited to the national border (of the United Kingdom, which includes today’s Republic of Ireland). Secondly, we analyse whether these provincial titles emphasised local events or increasingly attended to metropolitan news. Thirdly, we have a closer look at transnational aspects, more precisely the presence of popular imperialism in these working-class newspapers.

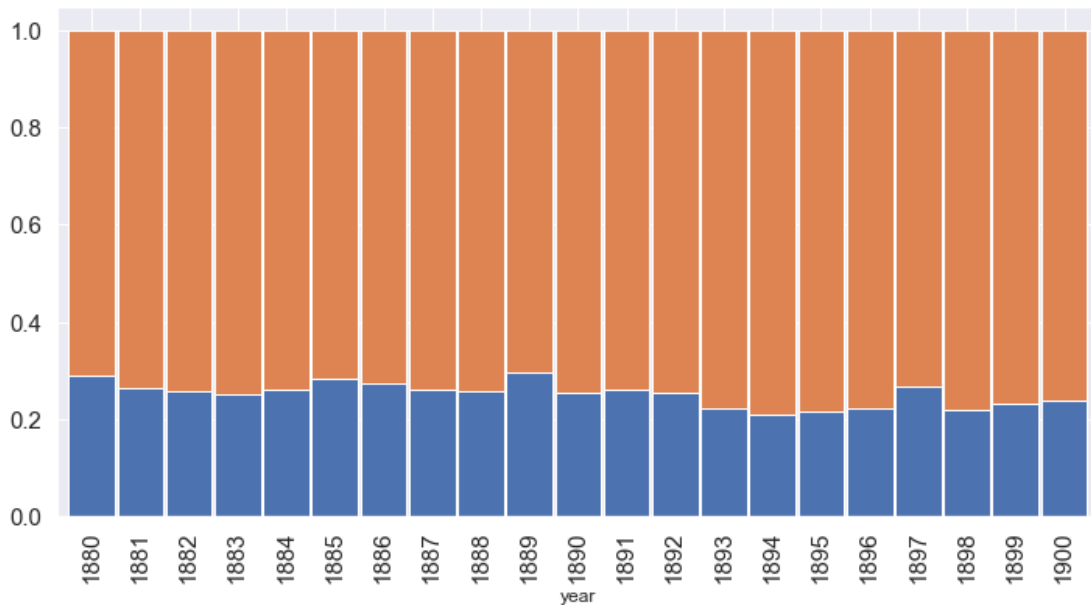


Figure 2: The proportion of places within the United Kingdom (orange bar) and outside its borders (blue bar).

Turning to the first question, Figure 2 shows the proportion of British (the orange bar) versus non-British places (the blue bar) between 1880 and 1900. While there is no dramatic change over these two decades, it shows a decrease in foreign place names (admittedly, a small drop), from 25% in the early 1880s to 23% around 1900. Put differently, more than 75% of all mentions comprise British place names. Taken together, these results may be more remarkable in their stability: while news reporting is often driven by unexpected events, on average, attention to what is happening outside the borders of the United Kingdom remained more or less unchanged.

Newspapers played a critical role in shaping and upholding the nation as an imagined community [2]. But, even though the previous analysis shows that the press was, discursively, firmly anchored in British soil, this doesn’t necessarily imply that it was “national” in its scope. In his extensive study *A Fleet Street in Every Town*, the historian Andrew Hobbs concedes that the Victorian reader generally preferred local news and that the press played a critical role in forging local communities [20]. Nonetheless, while the idea of a “national press” was still only emergent, these provincial papers were far from isolated entities. Hobbs, at the same

time, underlined the networked character of the Victorian provincial press. While newspapers often served as chroniclers of local culture and events, they did so as local nodes in a wider, national network of information. Put differently, the provincial press was “a ‘national’ [network] made from many ‘local’ elements”, and London figured as a central node in this network. This suggests that the provincial press was far from being parochial.

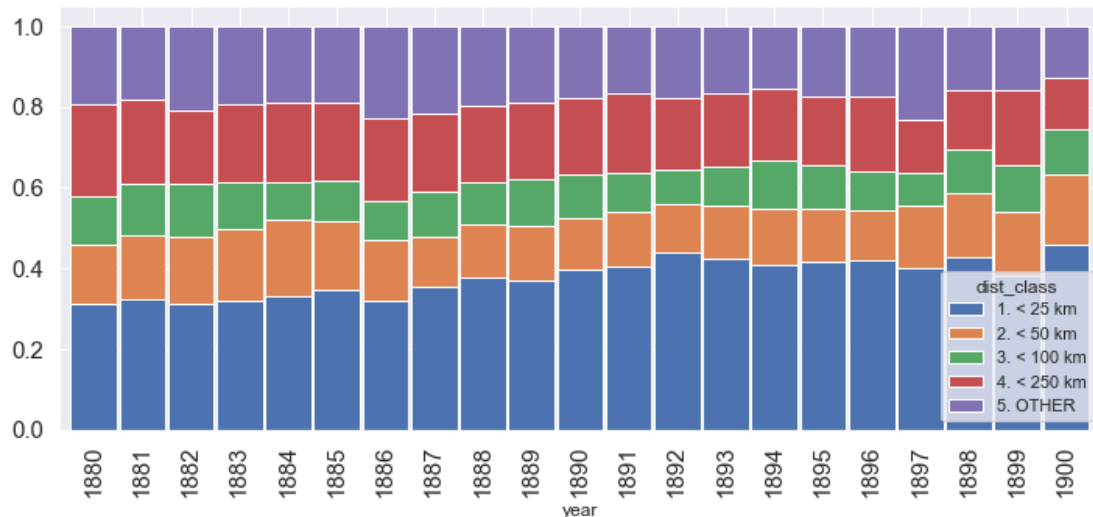


Figure 3: Toponyms by their distance to the newspaper’s places of publication.

To better understand how these newspapers meandered between the local and national level, we scrutinise the distribution of toponyms situated in England, Wales, Scotland and Ireland. For Figure 3, we first computed the distance between each toponym and the newspaper’s place of publication.²⁷ We divided these mentions into different bands based on their proximity to the place of publication. For example, the blue band shows the proportion of places names which were less than 25km removed from the district where the paper was produced. Interestingly enough, the geographical coverage of these papers tended to shrink. Changes remain small, but we do observe an increasing emphasis on more local matters (again, very rudimentarily measured as events taking place near the place of publication). On average, the proportion of toponyms in the blue band increases by roughly 5 percentage points over these two decades. To assess the dominance of the metropolis in the provincial press, we calculated the distance of each toponym to London.²⁸ While London was very present indeed, it was not as dominant as expected: less than 20% of all the toponyms were located in or around London. Most surprisingly, the number of mentions seems to decline over time, which ties in with our earlier finding that suggested a narrowing of the geographic horizon of these late Victorian titles.

²⁷We used historical press directories to determine the place of publication. For more information on the press directories, see [4].

²⁸We looked at places less than 25km removed from the coordinates as reported on Wikidata (<https://www.wikidata.org/wiki/Q84>).

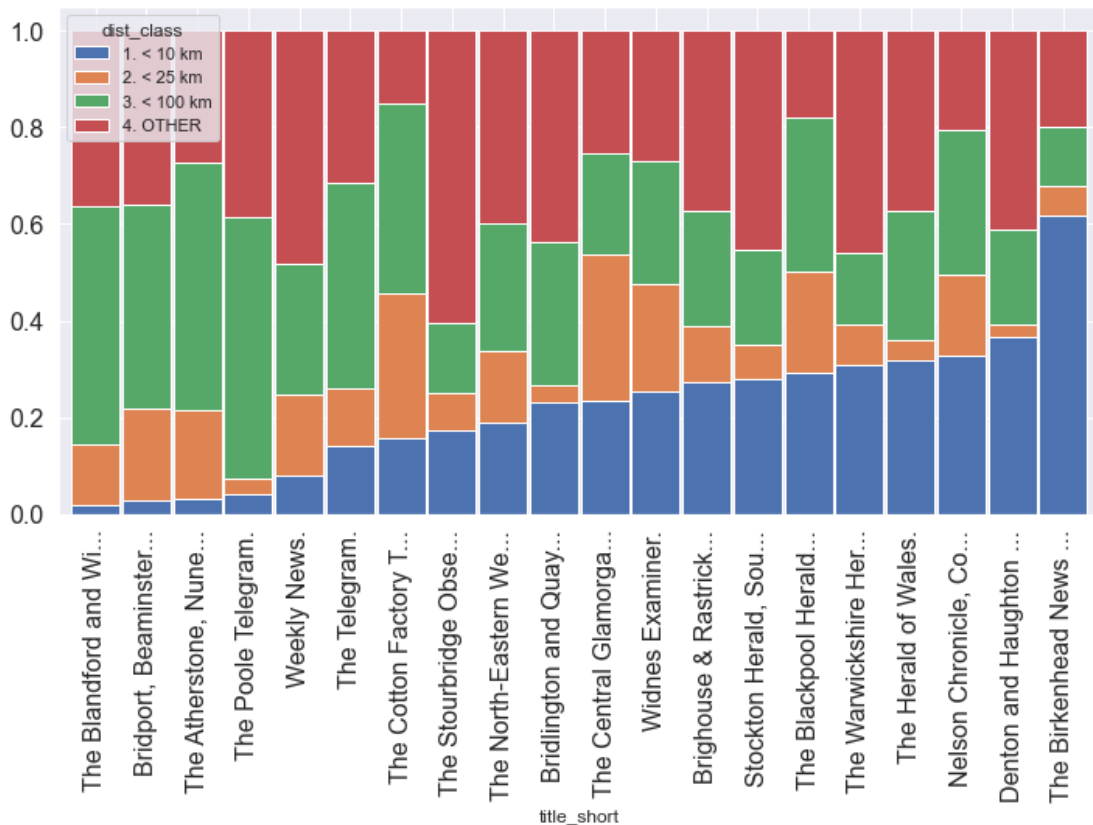


Figure 4: Toponyms by their distance to the newspaper's places of publication.

When comparing how individual papers vary in their emphasis on the local event, the differences become more pronounced, as shown in Figure 4. This allows us to understand and classify how these titles differed in terms of their geographical reach and coverage. Some of these provincial periodicals had a distinctively local emphasis. For example, close to 60% of all places in the *The Birkenhead News and Wirral General Advertiser* are located within a radius of 25km from Birkenhead. Others are broader in their coverage, they appear as less centred on one specific locality, but on a wider region. The *The Atherstone, Nuneaton, and Warwickshire Times* can serve as an example. Even though 50% of all toponyms are less than 50km removed from Atherstone, just about 20% appear within a 25km radius of this town. Exploring the distribution of these toponyms, therefore, might be a valuable way of understanding how these working-class papers anchored themselves spatially, negotiating between local, regional and national identities.

Lastly, we scrutinised the transnational level, focusing on the imperial geographies embedded in these digitised newspapers. In his analysis of popular imperialism, Nicholson [37] relies on popular provincial newspapers to probe the attitudes of the working classes towards the imperialist project. Especially concerning the Second Boer War (1899-1902), he questions

whether the working-class patriotic support for this endeavour was as strong as historians previously imagined. Looking at the results gathered from our corpus, it firstly transpires that geographical mentions of the empire were relatively low, consistently hovering around 5% of all the toponyms. Zooming in on Africa and Asia, the numbers are even lower—especially compared to references to locations in Canada and Australia—except around moments of crisis, such as the Second Boer War. Mentions of South African place names, for example, showed a dramatic increase at the end of the 19th century. These results are preliminary and should be complemented with additional content- and sentiment-based analyses in order to monitor more accurately the prevalence of jingoism in the popular press.

7. Conclusion

In this paper, we presented a comprehensive step-by-step examination of toponym linking for historical newspapers in English. We argued that a good performance on standard and highly generic benchmarks does not necessarily extrapolate to other domains. When applied to digitised historical newspapers, the accuracy of these state-of-the-art tools often drop significantly, therefore hinting at the complexity of finding a general solution to EL. We have presented and evaluated a new and very adaptable tool, T-Res, that resulted from these investigations: T-Res builds on top of robust NLP approaches, tailoring them to the specific task of toponym linking in historical newspapers. We concluded with a historical case study that demonstrated how our pipeline supports ongoing research on the local, national and transnational dimensions of the popular press.

Acknowledgements

The authors are grateful to the reviewers for their careful and constructive reviews. Work for this paper was produced as part of Living with Machines. This project, funded by the UK Research and Innovation (UKRI) Strategic Priority Fund, is a multidisciplinary collaboration delivered by the Arts and Humanities Research Council (AHRC grant AH/S01179X/1), with The Alan Turing Institute, the British Library and the Universities of Cambridge, East Anglia, Exeter, and Queen Mary University of London. This work was also supported by The Alan Turing Institute (EPSRC grant EP/N510129/1).

References

- [1] A. Akbik, D. Blythe, and R. Vollgraf. “Contextual string embeddings for sequence labeling”. In: *Proceedings of the 27th international conference on computational linguistics*. Santa Fe: Association for Computational Linguistics, 2018, pp. 1638–1649.
- [2] B. Anderson. *Imagined communities: Reflections on the origin and spread of nationalism*. Verso books, 2006.

- [3] M. Beals and E. Bell. “The atlas of digitised newspapers and metadata: Reports from Oceanic Exchanges”. In: *Loughborough: Loughborough University* (2020). DOI: 10.6084/m9.figshare.11560059.
- [4] K. Beelen, J. Lawrence, D. C. Wilson, and D. Beavan. “Bias and representativeness in digitized newspaper collections: Introducing the environmental scan”. In: *Digital Scholarship in the Humanities* 38.1 (2023), pp. 1–22. DOI: 10.1093/llc/fqac037.
- [5] E. Boros, C.-E. González-Gallardo, E. Giamphy, A. Hamdi, J. G. Moreno, and A. Doucet. “Knowledge-based contexts for historical named entity recognition & linking”. In: *Conference and Labs of the Evaluation Forum (CLEF)*. Vol. 3180. CEUR Workshop Proceedings, 2022.
- [6] E. Boros, E. L. Pontes, L. A. Cabrera-Diego, A. Hamdi, J. G. Moreno, N. Sidère, and A. Doucet. “Robust named entity recognition and linking on historical multilingual documents”. In: *Conference and Labs of the Evaluation Forum (CLEF)*. Vol. 2696. CEUR-WS Working Notes. 2020.
- [7] E. Boros, A. Hamdi, E. L. Pontes, L.-A. Cabrera-Diego, J. G. Moreno, N. Sidere, and A. Doucet. “Alleviating digitization errors in named entity recognition for historical documents”. In: *Proceedings of the 24th conference on computational natural language learning*. Acl, 2020, pp. 431–441. DOI: 10.18653/v1/2020.conll-1.35.
- [8] R. Bunescu and M. Paşca. “Using Encyclopedic Knowledge for Named entity Disambiguation”. In: *11th Conference of the European Chapter of the Association for Computational Linguistics*. Trento: Association for Computational Linguistics, 2006, pp. 9–16.
- [9] M. Coll Ardanuy, D. Beavan, K. Beelen, K. Hosseini, J. Lawrence, K. McDonough, F. Nanni, D. van Strien, and D. C. Wilson. “A dataset for toponym resolution in nineteenth-Century English newspapers”. In: *Journal of Open Humanities Data* 8 (2022). DOI: 10.5334/johd.56.
- [10] M. Coll Ardanuy, K. Hosseini, K. McDonough, A. Krause, D. van Strien, and F. Nanni. “A deep learning approach to geographical candidate selection through toponym matching”. In: *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*. 2020, pp. 385–388. DOI: 10.1145/3397536.3422236.
- [11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. “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)*. Minneapolis: Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.
- [12] M. L. Díez Platas, S. Ros Munoz, E. González-Blanco, P. Ruiz Fabo, and E. Alvarez Melado. “Medieval Spanish (12th–15th centuries) named entity recognition and attribute annotation system based on contextual information”. In: *Journal of the Association for Information Science and Technology* 72.2 (2021), pp. 224–238. DOI: 10.1002/asi.24399.
- [13] M. Ehrmann, G. Colavizza, Y. Rochat, and F. Kaplan. “Diachronic evaluation of NER systems on old newspapers”. In: *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*. Bochumer Linguistische Arbeitsberichte, 2016, pp. 97–107.

- [14] M. Ehrmann, M. Romanello, A. Flückiger, and S. Clemenide. “Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers”. In: *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*. Ed. by L. Cappellato, C. Eickhoff, N. Ferro, and A. Névél. Vol. 2696. Thessaloniki: Ceur-ws, 2020.
- [15] M. Ehrmann, M. Romanello, S. Najem-Meyer, A. Doucet, and S. Clemenide. “Overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents”. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF)*. Springer, 2022, pp. 423–446. DOI: 10.1007/978-3-031-13643-6_26.
- [16] P. Ferragina and U. Scaiella. “Tagme: on-the-fly annotation of short text fragments (by Wikipedia entities)”. In: *Proceedings of the 19th ACM international conference on Information and knowledge management*. New York: Association for Computing Machinery, 2010, pp. 1625–1628. DOI: 10.1145/1871437.1871689.
- [17] O.-E. Ganea and T. Hofmann. “Deep Joint Entity Disambiguation with Local Neural Attention”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen: Association for Computational Linguistics, 2017, pp. 2619–2629. DOI: 10.18653/v1/D17-1277.
- [18] C.-E. González-Gallardo, E. Boros, N. Girdhar, A. Hamdi, J. G. Moreno, and A. Doucet. “Yes but... Can ChatGPT identify entities in historical documents?” In: *arXiv preprint arXiv:2303.17322* (2023).
- [19] A. Hamdi, A. Jean-Caurant, N. Sidère, M. Coustaty, and A. Doucet. “Assessing and minimizing the impact of OCR quality on named entity recognition”. In: *Digital Libraries for Open Knowledge (TPDL)*. Springer, 2020, pp. 87–101. DOI: 10.1007/978-3-030-54956-5_7.
- [20] A. Hobbs. *A Fleet Street in every town: The provincial press in England, 1855-1900*. Open Book Publishers, 2018. DOI: 10.11647/obp.0152.
- [21] A. Hobbs. “The deleterious dominance of The Times in nineteenth-century scholarship”. In: *Journal of Victorian Culture* 18.4 (2013), pp. 472–497. DOI: 10.1080/13555502.2013.854519.
- [22] K. Hosseini, K. Beelen, G. Colavizza, and M. Coll Ardanuy. “Neural language models for nineteenth-century English”. In: *Journal of Open Humanities Data* (2021). DOI: 10.5334/johd.48.
- [23] K. Hosseini, F. Nanni, and M. Coll Ardanuy. “DeezyMatch: A flexible deep learning approach to fuzzy string matching”. In: *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations*. Online: Association for Computational Linguistics, 2020, pp. 62–69. DOI: 10.18653/v1/2020.emnlp-demos.9.
- [24] J. M. van Hulst, F. Hasibi, K. Dercksen, K. Balog, and A. P. de Vries. “REL: An Entity Linker Standing on the Shoulders of Giants”. In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Sigir ’20. New York: Acm, 2020, pp. 2197–2200. DOI: 10.1145/3397271.3401416.
- [25] N. Kolitsas, O.-E. Ganea, and T. Hofmann. “End-to-End Neural Entity Linking”. In: *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Brussels, 2018, pp. 519–529. DOI: 10.18653/v1/K18-1050.

- [26] K. Labusch and C. Neudecker. “Named Entity Disambiguation and Linking Historic Newspaper OCR with BERT”. In: *Conference and Labs of the Evaluation Forum (CLEF)*. Vol. 2696. CEUR Workshop Proceedings, 2020.
- [27] P. Le and I. Titov. “Improving entity linking by modeling latent relations between mentions”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne: Association for Computational Linguistics, 2018, pp. 1595–1604. DOI: 10.18653/v1/P18-1148.
- [28] J. L. Leidner. “Toponym resolution in text: annotation, evaluation and applications of spatial grounding”. In: *ACM SIGIR Forum*. Vol. 41. 2. New York: Association for Computing Machinery, 2007, pp. 124–126. DOI: 10.1145/1328964.1328989.
- [29] E. Linhares Pontes, L. A. Cabrera-Diego, J. G. Moreno, E. Boros, A. Hamdi, A. Doucet, N. Sidere, and M. Coustaty. “MELHISSA: a multilingual entity linking architecture for historical press articles”. In: *International Journal on Digital Libraries* 23.2 (2022), pp. 133–160. DOI: 10.1007/s00799-021-00319-6.
- [30] E. Linhares Pontes, A. Hamdi, N. Sidere, and A. Doucet. “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*. Springer, 2019, pp. 102–115. DOI: 10.1007/978-3-030-34058-2_11.
- [31] E. Manjavacas and L. Fonteyn. “Adapting vs. Pre-training Language Models for Historical Languages”. In: *Journal of Data Mining & Digital Humanities Nlp4dh* (2022). DOI: 10.46298/jdmdh.9152.
- [32] K. McDonough, L. Moncla, and M. van de Camp. “Named entity recognition goes to old regime France: geographic text analysis for early modern French corpora”. In: *International Journal of Geographical Information Science* 33.12 (2019), pp. 2498–2522. DOI: 10.1080/13658816.2019.1620235.
- [33] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. “DBpedia spotlight: shedding light on the web of documents”. In: *Proceedings of the 7th international conference on semantic systems*. 2011, pp. 1–8. DOI: 10.1145/2063518.2063519.
- [34] R. Mihalcea and A. Csomai. “Wikify! Linking documents to encyclopedic knowledge”. In: *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. New York: Association for Computing Machinery, 2007, pp. 233–242. DOI: 10.1145/1321440.1321475.
- [35] D. Milne and I. H. Witten. “Learning to link with Wikipedia”. In: *Proceedings of the 17th ACM conference on Information and knowledge management*. New York: Association for Computing Machinery, 2008, pp. 509–518. DOI: 10.1145/1458082.1458150.
- [36] G. Munnely and S. Lawless. “Investigating entity linking in early English legal documents”. In: *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*. New York: Acm, 2018, pp. 59–68. DOI: 10.1145/3197026.3197055.
- [37] J. Nicholson. “Popular Imperialism and the Provincial Press: Manchester Evening and Weekly Papers, 1895-1902”. In: *Victorian Periodicals Review* 13.3 (1980), pp. 85–96.

- [38] A. Olieman, K. Beelen, M. van Lange, J. Kamps, and M. Marx. “Good Applications for Crummy Entity Linkers? The Case of Corpus Selection in Digital Humanities”. In: *arXiv preprint arXiv:1708.01162*. 2017.
- [39] N. Pedrazzini and B. McGillivray. “Machines in the media: semantic change in the lexicon of mechanization in 19th-century British newspapers”. In: *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*. Taipei: Association for Computational Linguistics, 2022, pp. 85–95.
- [40] F. Piccinno and P. Ferragina. “From TagME to WAT: a new entity annotator”. In: *Proceedings of the first international workshop on Entity recognition & disambiguation*. New York: Association for Computing Machinery, 2014, pp. 55–62. DOI: 10.1145/2633211.2634350.
- [41] L. Ratinov, D. Roth, D. Downey, and M. Anderson. “Local and global algorithms for disambiguation to Wikipedia”. In: *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*. Portland, 2011, pp. 1375–1384.
- [42] M. Rovera, F. Nanni, S. P. Ponzetto, and A. Goy. “Domain-specific Named Entity Disambiguation in Historical Memoirs”. In: *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*. Vol. 2006. CEUR Workshop Proceedings. Rome, 2017.
- [43] R. Santos, P. Murrieta-Flores, P. Calado, and B. Martins. “Toponym matching through deep neural networks”. In: *International Journal of Geographical Information Science* 32.2 (2018), pp. 324–348. DOI: 10.1080/13658816.2017.1390119.
- [44] S. Schweter, L. März, K. Schmid, and E. Çano. “hmbERT: Historical Multilingual Language Models for Named Entity Recognition”. In: *Conference and Labs of the Evaluation Forum (CLEF)*. Vol. 3180. CEUR Workshop Proceedings, 2022.
- [45] A. Sil, G. Kundu, R. Florian, and W. Hamza. “Neural cross-lingual entity linking”. In: *Thirty-Second AAAI Conference on Artificial Intelligence*. AAAI Press, 2018, pp. 5464–5472.
- [46] D. van Strien, K. Beelen, M. Coll Ardanuy, K. Hosseini, B. McGillivray, and G. Colavizza. “Assessing the impact of OCR quality on downstream NLP tasks”. In: *Proceedings of the 12th International Conference on Agents and Artificial Intelligence (ICAART)*. Volume 1: ARTIDIGH. 2020, pp. 484–496.
- [47] G. Tolfo, O. Vane, K. Beelen, K. Hosseini, J. Lawrence, D. Beavan, and K. McDonough. “Hunting for Treasure: Living with Machines and the British Library Newspaper Collection”. In: *Digitised Newspapers – A New Eldorado for Historians?: Reflections on Tools, Methods and Epistemology*. Ed. by E. Bunout, M. Ehrmann, and F. Clavert. De Gruyter Oldenbourg, 2023, pp. 23–46. DOI: 10.1515/9783110729214-002.
- [48] M. Van Erp, P. Mendes, H. Paulheim, F. Ilievski, J. Plu, G. Rizzo, and J. Waitelonis. “Evaluating entity linking: An analysis of current benchmark datasets and a roadmap for doing a better job”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. European Language Resources Association, 2016, pp. 4373–4379.

- [49] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. “Transformers: State-of-the-Art Natural Language Processing”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, 2020, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6.
- [50] I. Yamada, A. Asai, J. Sakuma, H. Shindo, H. Takeda, Y. Takefuji, and Y. Matsumoto. “Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, 2020, pp. 23–30. DOI: 10.18653/v1/2020.emnlp-demos.4.
- [51] I. Yamada, H. Takeda, and Y. Takefuji. “Enhancing named entity recognition in Twitter messages using entity linking”. In: *Proceedings of the Workshop on Noisy User-generated Text*. Beijing: Association for Computational Linguistics, 2015, pp. 136–140. DOI: 10.18653/v1/W15-4320.

A. Appendix: T-Res

The following code snippet shows how the T-Res pipeline works:

```

from geoparser import pipeline, recogniser, ranking, linking

myner = recogniser.Recogniser(...) # Instantiate the Recogniser
myranker = ranking.Ranker(...) # Instantiate the Ranker
mylinker = linking.Linker(...) # Instantiate the Linker

geoparser = pipeline.Pipeline(myner=myner, myranker=myranker, mylinker=mylinker)

output = geoparser.run_text("Inspector Liddle said: I am an inspector of police, living
                             in the city of Durham.",
                             place="Alston, Cumbria, England",
                             place_wqid="Q2560190"
                             )

```

The parentheses (...) indicate an ellipsis in the code, where the user has the option to instantiate each of the three modules (the Recogniser for named entity recognition, the Ranker for candidate selection, and the Linker for entity disambiguation) according to their needs. For example, they may choose to instantiate a Recogniser that uses a specific model for named entity recognition from the HuggingFace hub, or they may choose to train their own model, provided a base model and their own dataset (in the format required). They may instantiate a Ranker that, given a KB, uses the *exact* match approach to find candidates, or choose to train their own DeezyMatch model, given a dataset of positive and negative pairs, and use it for candidate selection. Likewise, they may instantiate a Linker module that, given a KB, uses the *mostpopular* approach, or they may train their own *rel* disambiguation approach.

The following snippet shows the output from the previous command, as a json:

```
[{"mention": "Durham",
```

```

"ner_score": 0.999,
"pos": 74,
"sent_idx": 0,
"end_pos": 80,
"tag": "LOC",
"sentence": "Inspector Liddle said: I am an inspector of police, living
             in the city of Durham.",
"prediction": "Q179815",
"ed_score": 0.039,
"cross_cand_score": {
  "Q179815": 0.396,
  "Q23082": 0.327,
  "Q49229": 0.141,
  "Q5316459": 0.049,
  "Q458393": 0.045,
  "Q17003433": 0.042,
  "Q1075483": 0.0
},
"string_match_score": {"Durham": [1.0, ["Q1137286", "Q5316477", "Q752266", "..."]]},
"prior_cand_score": {
  "Q179815": 0.881,
  "Q49229": 0.522,
  "Q5316459": 0.457,
  "Q17003433": 0.455,
  "Q23082": 0.313,
  "Q458393": 0.295,
  "Q1075483": 0.293
},
"latlon": [54.783333, -1.566667],
"wkd_t_class": "Q515"}

```

For each mention detected in the input text, our tool returns:

- mention: mention as it appears in the text.
- ner_score: NER confidence score.
- pos: start character position of the mention in the sentence.
- sent_idx: sentence index in the text.
- end_pos: end character position of the mention in the sentence.
- tag: name entity type.
- sentence: target sentence.
- prediction: predicted Wikidata entity.
- ed_score: disambiguation confidence score.
- cross_cand_score: selected candidates and their cross-candidate confidence scores.
- string_match_score: selected candidates and their string matching confidence scores.
- prior_cand_score: selected candidates and their prior confidence scores.
- latlon: geographic coordinates of the predicted entity.
- wkd_t_class: Most common Wikidata class of the predicted entity.

The tool can also be used in a step-wise manner, or for just one module in the pipeline. We provide full documentation in our GitHub repository: <https://github.com/Living-with-machines/T-Res>.