Enriching the Metadata of Community-Generated Digital Content through Entity Linking: An Evaluative Comparison of State-of-the-Art Models

Youcef Benkhedda[*,1], **Adrians Skapars**[*,1], **Viktor Schlegel**[1,2],
Goran Nenadic[1] and **Riza Batista-Navarro**[1]
[1]Department of Computer Science, University of Manchester, UK
[2]ASUS Intelligent Cloud Services (AICS), Singapore
youcef.benkhedda@manchester.ac.uk, adrians.skapars@postgrad.manchester.ac.uk,
viktor_schlegel@asus.com, {gnenadic,riza.batista}@manchester.ac.uk

Abstract

Digital archive collections that have been contributed by communities, known as community-generated digital content (CGDC), are important sources of historical and cultural knowledge. However, CGDC items are not easily searchable due to semantic information being obscured within their textual metadata. In this paper, we investigate the extent to which state-of-the-art, general-domain entity linking (EL) models (i.e., BLINK, EPGEL and mGENRE) can map named entities mentioned in CGDC textual metadata, to Wikidata entities. We evaluate and compare their performance on an annotated dataset of CGDC textual metadata and provide some error analysis, in the way of informing future studies aimed at enriching CGDC metadata using entity linking methods.

1 Introduction

Community-generated digital content (CGDC) pertains to digital-born archive collections that have been developed by communities. In the UK, for instance, libraries and museums such as the Morrab Library in Cornwall[1] and the Sherborne Museum in Dorset[2] employ volunteers to catalogue their archive collections, consisting of historic photographs and papers, respectively. Since 1994, the UK National Lottery Heritage Fund has awarded grants to around 5000 community history projects,[3] leading to the proliferation of CGDC.

Encapsulating the collective experiences and narratives of communities over time, such collections serve as indispensable sources of knowledge, offering a window into the past that deepens our understanding of human history and culture (Konstantelos et al., 2019). However, despite their important role in enhancing people's appreciation of their heritage, CGDC items remain hard to find (Hanna et al., 2021). This can be attributed to the fact that semantic information on CGDC items tends to be buried within their textual metadata, e.g., titles and descriptions, making it difficult to search for and link items related to a given query. Such a challenge can be potentially overcome by enriching CGDC metadata using natural language processing (NLP) methods. As a first step, for example, named entity recognition (NER) can be employed to automatically label the names of any entities mentioned within a piece of text (Humbel et al., 2021; Jehangir et al., 2023). This is often followed by entity linking (EL), a task concerned with normalising name variants (e.g., the canonical and vernacular names of a place) to the same real-world entity (Oliveira et al., 2021); typically, this is implemented as a disambiguation task where a unique identifier (denoting an entity) used within a knowledge base, is assigned to a given named entity.

In this paper, we focus on assessing the performance of state-of-the-art EL models on CGDC textual metadata. These models have demonstrated impressive EL performance in the general domain, e.g., on the task of Wikification which involves linking entities within text to Wikipedia (Moro et al., 2014). CGDC metadata, however, are not as well-formed as general-domain texts such as news articles, mainly due to the fact that there are no established standards that require communities to write their textual metadata in a consistent way. For instance, many CGDC descriptions are short, consisting only of phrases rather than full sentences; misspellings and obsolete names are also commonplace. We thus aim to evaluate how well existing state-of-the-art models perform on CGDC textual metadata and analyse cases on which these models tend to fail. This will help researchers working in the areas of digital humanities and cultural analytics in identifying ways on how existing EL approaches can be adapted or optimised for CGDC.

[*]These authors contributed equally to this work.
[1]https://morrablibrary.org.uk/
[2]https://www.sherbornemuseum.com/
[3]https://www.heritagefund.org.uk/our-work/museums-libraries-and-archives

Proceedings of LaTeCH-CLfL 2024, pages 213–220

To the best of our knowledge, ours is the first work to explore EL for CGDC. The handful of efforts that employed EL on archive collections focussed mostly on historical newspapers (Labusch and Neudecker, 2020; Ehrmann et al., 2020; Linhares Pontes et al., 2022; Hamdi et al., 2021), specific centuries (Brando et al., 2015, 2016) or events such as the Second World War (Heino et al., 2017), but not on CGDC.

2 Dataset

To support our evaluation of state-of-the-art entity linking models, we set out to develop our own annotated dataset of CGDC textual metadata.

2.1 Data Collection

We collected textual metadata written in English for items in the following CGDC archives:

Spratton Local History Society Collection (Spratton). Based in the village of Spratton, Northamptonshire in the UK, the Spratton Local History Society[4] have created web pages containing short biographies of Spratton men who served in the First World War.

National Lottery Heritage Fund (NLHF) Archives. Various community projects that were given grants by the UK NLHF have created web pages documenting the lives of people relevant to the history of the communities. These include: Vale People First,[5] The Friends of Hemingfield Colliery,[6] Dorset Ancestors,[7] Farnhill World War I Volunteers[8] and The Haringey First World War Peace Forum.[9]

The Morrab Library Photographic Archive (Morrab). This archive[10] contains over 15,000 digitised photographs capturing Cornish history and culture. Each photograph comes with textual metadata such as title and description.

People's Collection Wales (PCW). PCW[11] is an online platform that allows individuals, community groups and small museums/libraries to contribute items pertaining to Welsh culture and history, including photographs, documents, and audio and video recordings. For each of the more than 150K items in PCW, a title and a description are provided.

We sampled 20 items from the Spratton collection, 25 from the NLHF archives, 50 from the Morrab collection and 50 from PCW. Based on these items, we created the documents that comprise our CGDC dataset. In the case of the Spratton and NLHF subsets, each document contains the title and full text of a web page. Meanwhile, each document in the Morrab and PCW subsets consists of the concatenation of the title and description of the corresponding item.

2.2 Data Annotation

The documents in our collected data were labelled according to the two types of annotations described below, with the help of the brat[12] rapid annotation tool (Stenetorp et al., 2012).

Annotation of Named Entities. The span and semantic type of any named entity that falls under any of the following types were annotated: `Person` (`Per`), `Organisation` (`Org`), `Location` (`Loc`), `Miscellaneous` (`Misc`) and `Date`.

Annotation of Entity Links. All annotated named entities (except those that were given the `Date` label[13]) were linked to their unique identifiers in Wikidata.[14] If an entity cannot be found in Wikidata, it was linked to the NIL entity.

Guidelines (an overview of which is provided in Appendix A) were prepared, outlining details of the annotation task. Following these guidelines, one annotator labelled all 145 documents in our CGDC dataset. To allow for assessment of reliability of their annotations, a second annotator independently labelled a subset of 20 documents.

Based on the work of the two annotators, we measured inter-annotator agreement (IAA) for each of the two annotation types. When it comes to the annotation of named entities, an IAA of 74.8% in terms of F1-score was obtained. Taking only the named entities whose spans were labelled in the same way by both annotators, we measured IAA with respect to the annotation of entity links. The IAA in terms of Cohen's Kappa (Cohen, 1960) is

[4] http://www.sprattonhistory.org/
[5] https://www.valepeoplefirst.org.uk/dejavu/
[6] https://hemingfieldcolliery.org/people-lives-and-loss/
[7] https://dorset-ancestors.com/
[8] http://www.farnhill.co.uk/volunteers/articles/articles-people/
[9] https://hfwwpf.wordpress.com/
[10] https://photoarchive.morrablibrary.org.uk/
[11] https://www.peoplescollection.wales/
[12] https://brat.nlplab.org/
[13] In our study, temporal expressions are not considered to be real-world entities that need to be linked to Wikidata.
[14] https://www.wikidata.org/wiki/Wikidata:Main_Page

77.79%, which is considered to be substantial (Landis and Koch, 1977). The labels provided by the second author of this paper serve as gold standard annotations.

The 45 annotated documents in the Spratton and NLHF subsets were held out and were used to identify the values of parameters that need to be configured to run the EL models that we selected for comparison (described in the next section). Meanwhile, the annotated documents in the Morrab and PCW subsets (100 overall) were considered as test data, forming the basis of the evaluation of the performance of the chosen EL models.[15] Table 3 in Appendix B summarises the number of named entities per type in the said test data.

3 Problem Formulation and Models

We first provide a formal definition of the EL task: given a target knowledge base containing a set of entities E and a textual document in which a set of named entities N have been identified, an EL model maps each $n \in N$ to the corresponding entity $e \in E$ in the knowledge base. If the entity that n corresponds to does not exist in E, then n is considered to be unlinkable and is thus linked to a NIL entity. In our work, the context in which n appears is also provided as input (together with n itself) and the target knowledge base is Wikidata.

Three state-of-the-art EL models were investigated in this study. For a given named entity (NE), each of the models predicts the best matching entity in the knowledge base that it should be linked to by specifying its identifier (ID) together with a similarity score, if it is linkable; otherwise, it is linked to NIL.

BLINK. This model (Wu et al., 2020) employs BERT-based architectures (Devlin et al., 2019) for two subtasks: retrieving candidate entities by encoding the context containing an NE and the definitions of candidate entities in Wikipedia, and ranking the candidates. Its predicted Wikipedia IDs are mapped to Wikidata IDs.

Entity Profile Generation for Entity Linking (EPGEL). This model (Lai, 2022) makes use of the BART sequence-to-sequence model (Lewis et al., 2020) and a dictionary-based method to generate a "profile", i.e., a title and description, for a given NE based on the context in which it appears. These profiles are then used to retrieve candidate matching entities within Wikidata.

Multilingual Generative Entity Retrieval (mGENRE). Unlike the two models above, mGENRE (De Cao et al., 2022) is capable of linking named entities to a multilingual knowledge base. It employs a pre-trained multilingual BART model that takes an NE and auto-regressively generates its Wikipedia name, which is then mapped to the corresponding Wikidata ID.

In order to identify what configuration of the above models should be used in applying them on our CGDC test data, we utilised our held-out data to determine: (1) how much context should be provided as input to the model together with an NE, and (2) the threshold for the similarity score, whereby an NE is linked to NIL if the similarity score of its top-matching candidate is lower than the threshold. We observed that for all models, providing the sentence immediately preceding and succeeding the sentence containing a given NE, leads to optimal results. Meanwhile, the following values were found to be ideal similarity thresholds: 0.7, 0.8 and 0.4 for BLINK, EPGEL and mGENRE, respectively.

4 Results and Discussion

The models were applied to the gold standard `Person`, `Organisation`, `Location` and `Miscellaneous` NEs in our CGDC test set, i.e., the Morrab and PCW subsets. It is worth noting that we utilised each of the chosen EL models out-of-the-box, i.e., without extending their functionality. All models make use of the text span of a given NE in their analysis, but none of them consider the NE type as a feature, although it is available as part of the input to EL.

A preliminary check was performed to detect `Person` NEs that contain only one token; such NEs were automatically given NIL as their ID, as our preliminary experimentation with the held-out dataset showed that the three EL models are unlikely to be able to correctly disambiguate them.

4.1 Evaluation Metrics

EL performance is typically evaluated in terms of accuracy, i.e., the number of correctly linked NEs over the total number of NEs in the evaluation data. Taking inspiration from the work of Zhu et al.

[15] Our annotations, provided in the standoff format supported by the brat tool, are available for download at https://github.com/OurHeritageOurStories/cgdc_annotations.

NE Type	Model	Non-NAC	NAC	OAC
Per	BLINK	0.412	0.782	0.735
	EPGEL	**0.588**	0.681	0.669
	mGENRE	0.389	**0.925**	**0.855**
Org	BLINK	0.694	0.000	0.426
	EPGEL	**0.895**	0.444	0.720
	mGENRE	0.772	**0.946**	**0.840**
Loc	BLINK	0.808	0.000	0.652
	EPGEL	**0.795**	0.524	**0.743**
	mGENRE	0.708	**0.746**	0.716
Misc	BLINK	**0.750**	0.000	0.488
	EPGEL	0.714	0.357	0.595
	mGENRE	0.607	**1.000**	**0.738**
ALL	BLINK	0.767	0.392	0.621
	EPGEL	**0.795**	0.582	0.712
	mGENRE	0.695	**0.885**	**0.769**

Table 1: EL results on the test data according to named entity (NE) type. Key: Non-NAC = non-NIL accuracy; NAC = NIL accuracy; OAC = overall accuracy.

(2023), we report the performance of the three EL models according to three types of accuracy: (1) non-NIL accuracy (Non-NAC), which considers only NEs that are linked to Wikidata IDs according to the gold standard; (2) NIL accuracy, which considers only unlinkable NEs, i.e., those that are linked to NIL, according to the gold standard; and (3) overall accuracy (OAC), which considers all NEs regardless of whether they are linked to Wikidata IDs according to the gold standard or not.

Table 1 presents the performance of each of the models for each NE type and for the entire CGDC test data (ALL). Overall, EPGEL is best at predicting the IDs of linkable NEs. However, mGENRE is much better at identifying unlinkable (NIL) named entities, which are quite common in CGDC collections as many entities described in such collections are known only to local communities and thus do not have Wikidata entries. This positively impacted overall accuracy, leading to mGENRE obtaining optimal performance on the entire test set. The same trend can be observed for every NE type, except for the `Loc` type, where EPGEL obtained the best overall performance. A similar observation can be made when considering the performance of the models on each of the CGDC subsets that comprise the test data (see Table 4 in Appendix C).

4.2 Error Analysis

In Table 2, we provide examples of NEs that were wrongly linked by any of the three EL models, highlighting cases within CGDC that led to erroneous predictions. Firstly, lesser-known NEs that coincidentally share names with other entities pose a challenge to all models. For instance, the unlinkable NE *"Constance Amelia Browne"* was linked to *"Q75857857: Constance Browne"*, a member of the British peerage, by EPGEL (Example 1); *"Sir James Jebusa Shannon"* was linked to a painting with a similar name, *"Q28051261: James Jebusa Shannon"*, by BLINK (Example 2). All models wrongly linked *"Morrison Road"* to a road in Australia with the same name (Example 3).

A case that was found difficult by mGENRE in particular is Example 4. It wrongly linked *"Duenna"*, a play, to *"Q1519901: chaperone"*, as *"dueña"* happens to be a synonym for *"chaperone"* in Spanish.

All models seem to struggle to correctly link obsolete names (i.e., names that entities were formerly known as). An example of this is *"County Club hotel"* which is the old name of the entity *"Q1045316: Kings Head Hotel"* (Example 5). Both BLINK and EPGEL linked it to the wrong entity, while mGENRE considered it to be unlinkable and thus linked it to NIL.

4.3 Potential Applications

In this work, we have demonstrated the linking of NEs within CGDC textual metadata to Wikidata, a centralised knowledge base containing structured information on entities and concepts. In this way, our approach has a strong potential to improve the searchability of CGDC items. In the current scenario in which CGDC is made available by communities, a historian might struggle to find CGDC items that are described in their metadata using a vernacular name of a place, for instance. Mapping entities to Wikidata (or any other relevant knowledge base) will enable finding of such items, as all name variants of the same entity will have been assigned the same identifier by an EL model.

Currently, CGDC items collected by different archives or communities are siloed: they are findable only within a community's own catalogue or archive, but not across different archives. Linking CGDC textual metadata to a central knowledge base (Wikidata) via EL will make it possible to create a knowledge graph whereby CGDC items are linked based on the identified entities that they contain. This, in turn, will support seamless searching for CGDC items that pertain to particular entities of interest. Ideally, such a knowledge graph would be editable, allowing for human-in-the-loop data curation whereby human contributors can correct any erroneous identifiers assigned by EL models.

	NE with Context	NE Type	Gold Std.	BLINK	EPGEL	mGENRE
1	**Constance Amelia Browne** was the maternal...	Per	NIL	NIL	Q75857857: Constance Browne	NIL
2	The Bathers/ **Sir James Jebusa Shannon**/ 1900-23...	Per	Q731056: James Jebusa Shannon	Q28051261: Sir James Jebusa Shannon	Q731056: James Jebusa Shannon	Q731056: James Jebusa Shannon
3	Sandfields Branch Library, **Morrison Road**,...	Loc	NIL	Q6914046: Morrison Road	Q6914046: Morrison Road	Q6914046: Morrison Road
4	...Theatre Playbill, Advertising the play **Duenna**, to be staged at...	Misc	Q7731154: The Duenna	Q7731154: The Duenna	NIL	Q1519901: chaperone
5	...outside the old **County Club hotel** (a hospital), now...	Loc	Q1045316: Kings Head Hotel	Q6772978: London Marriott Hotel County Hall	Q55782270: Club Hotel	NIL

Table 2: Example input named entities (in bold) and context for which any of the three EL models produced erroneous predictions (shown in grey). The full context for each example can be found in Appendix D.

5 Conclusion and Future Work

In this paper, we present the results of evaluating BLINK, EPGEL and mGENRE— three of the state-of-the-art, general-domain entity linking models— on an annotated dataset of CGDC textual metadata. Our evaluation shows that mGENRE obtains superior performance overall and on unlinkable (NIL) named entities more specifically, which tend to be prevalent in CGDC. Our future work will focus on handling cases that make CGDC textual metadata particularly challenging, e.g., lesser-known named entities and obsolete names, and on combining the strengths of EPGEL and mGENRE in predicting IDs for linkable and unlinkable named entities, respectively. Furthermore, we will investigate how the performance of these models can be enhanced by making them leverage the NE type of any given named entity as a semantic feature that can inform the EL process.

Acknowledgement

This work was supported by the Arts and Humanities Research Council [grant number AH/W00321X/1 – Our Heritage, Our Stories: Linking and searching community-generated digital content to develop the people's national collection].

References

Carmen Brando, Francesca Frontini, and Jean-Gabriel Ganascia. 2015. Disambiguation of named entities in cultural heritage texts using linked data sets. In *New Trends in Databases and Information Systems: ADBIS 2015 Short Papers and Workshops, BigDap, DCSA, GID, MEBIS, OAIS, SW4CH, WISARD, Poitiers, France, September 8-11, 2015. Proceedings*, pages 505–514. Springer.

Carmen Brando, Francesca Frontini, and Jean-Gabriel Ganascia. 2016. REDEN: named entity linking in digital literary editions using linked data sets. *Complex Systems Informatics and Modeling Quarterly*, (7):60–80.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2022. Multilingual autoregressive entity linking. *Transactions of the Association for Computational Linguistics*, 10:274–290.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020. Overview of CLEF HIPE 2020: Named entity recognition and linking on historical newspapers. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings 11*, pages 288–310. Springer.

Ahmed Hamdi, Elvys Linhares Pontes, Emanuela Boros, Thi Tuyet Hai Nguyen, Günter Hackl, Jose G Moreno, and Antoine Doucet. 2021. A multilingual dataset for named entity recognition, entity linking and stance detection in historical newspapers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2328–2334.

Emma Hanna, Lorna M. Hughes, Lucy Noakes, Catriona Pennell, and James Wallis. 2021. Reflections on

the Centenary of the First World War: Learning and Legacies for the Future. Technical report, Arts and Humanities Research Council (AHRC).

Erkki Heino, Minna Tamper, Eetu Mäkelä, Petri Leskinen, Esko Ikkala, Jouni Tuominen, Mikko Koho, and Eero Hyvönen. 2017. Named entity linking in a complex domain: Case second world war history. In *Language, Data, and Knowledge: First International Conference, LDK 2017, Galway, Ireland, June 19-20, 2017, Proceedings 1*, pages 120–133. Springer.

Marco Humbel, Julianne Nyhan, Andreas Vlachidis, Kim Sloan, and Alexandra Ortolja-Baird. 2021. Named-entity recognition for early modern textual documents: a review of capabilities and challenges with strategies for the future. *Journal of Documentation*, 77(6):1223–1247.

Basra Jehangir, Saravanan Radhakrishnan, and Rahul Agarwal. 2023. A survey on Named Entity Recognition — datasets, tools, and methodologies. *Natural Language Processing Journal*, 3:100017.

Leo Konstantelos, Lorna Hughes, and William Kilbride. 2019. The Bits Liveth Forever? Digital Preservation and the First World War Commemoration. Technical report, IWM War and Conflict Subject Network.

Kai Labusch and Clemens Neudecker. 2020. Named Entity Disambiguation and Linking Historic Newspaper OCR with BERT. In *CLEF (Working Notes)*.

Ngoc Lai. 2022. LMN at SemEval-2022 task 11: A transformer-based system for English named entity recognition. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1438–1443, Seattle, United States. Association for Computational Linguistics.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Jose G Moreno, Emanuela Boros, Ahmed Hamdi, Antoine Doucet, Nicolas Sidere, and Mickaël Coustaty. 2022. MELHISSA: a multilingual entity linking architecture for historical press articles. *International Journal on Digital Libraries*, pages 1–28.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Italo L Oliveira, Renato Fileto, René Speck, Luís PF Garcia, Diego Moussallem, and Jens Lehmann. 2021. Towards holistic entity linking: Survey and directions. *Information Systems*, 95:101624.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable Zero-shot Entity Linking with Dense Entity Retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

Fangwei Zhu, Jifan Yu, Hailong Jin, Juanzi Li, Lei Hou, and Zhifang Sui. 2023. Learn to Not Link: Exploring NIL Prediction in Entity Linking. In *Findings of the Association for Computational Linguistics (ACL 2023)*, pages 10846–10860, Toronto, Canada. Association for Computational Linguistics.

A Annotation Guidelines

A.1 Annotation of Named Entities

Any named entity that falls under any of the following entity types should be annotated: `Person`, `Organisation`, `Location`, `Miscellaneous` and `Date`.

`Person`: names of people, e.g., *"Mary"*, *"John Smith"*.

`Organisation`: names of companies, authorities, institutions, agencies, groups of people, e.g., *"Navy"*, *"Home Office"*.

`Location`: names of places, cities, towns, streets, e.g., *"Camden"*, *"Abbey Road"*. Notes:

- depending on the context, the name of a place might be used to refer to an organisation or geo-political entity rather than the place itself, e.g., *"Westminster"* in *"Westminster made the announcement."* In such cases, the name should be annotated as an `Organisation` rather than as a `Location`.
- if the text mentions an address, i.e., a street name immediately followed by its city, the street name and city names should be annotated separately, e.g., *"Princess St"* and *"Manchester"* instead of *"Princess St, Manchester"*.

`Date`: temporal expressions, including both specific and ambiguous mentions of time, e.g., *"December 1950"*, *"early 50s"*, *"previous year"*. If the expression pertains to a range, each constituent temporal expression should be annotated separately, e.g., *"1970"* and *"1980"* instead of *"1970-1980"*.

`Miscellaneous`: a catch-all category for named entities that do not fall under any of the above entity types and yet might be of interest to historians/researchers, e.g., names of warships (e.g., *"Aida Lauro"*), infrastructure (e.g., *"HS2"*), demonyms (e.g., *"French"*). Importantly, this category includes named events, e.g., *"World War II"*.

Handling nested entities. Some sentences might contain nested entities, i.e., an entity within another entity, e.g., *"London"* in *"London Bridge"*. In such cases, only the outer entity, e.g., *"London Bridge"*, should be annotated.

Handing discontinuous entities. Some sentences might contain discontinuous entities, i.e., an entity whose tokens do not appear in one contiguous text span, e.g., *Lord Eskrine* in *"Lord and Lady Eskrine"* and *Battle of Gaza* in *"Battle of Rafa, Gaza and Jerusalem"*. In such cases, the text span should be decomposed into its constituent entities, e.g., *"Lord"* and *"Lady Eskrine"* (`Person` entities); and *"Battle of Rafa"*, *"Gaza"* and *"Jerusalem"* (`Miscellaneous` entities). Note how *"Gaza"* and *"Jerusalem"* were labelled as `Miscellaneous` entities; this is because they were interpreted as pertaining to the *Battle of Gaza* and the *Battle of Jerusalem*, rather than just *"Gaza"* and *"Jerusalem"*.

Handling co-referring expressions. Although a co-referring expression (e.g., *"he"*, *"the company"*) might pertain to a named entity mentioned within the text, we are not annotating coreference in this task so such expressions should simply be ignored.

A.2 Annotation of Entity Links

All annotated named entities should be linked to their Wikidata identifier (by specifying the full URL to the identified item in Wikidata), with the exception of entities that were given the `Date` label. In determining the correct identifier, it is acceptable to make use of any information available within Wikidata, e.g., definitions, synonyms or properties of a candidate item. If an entity cannot be found in Wikidata, the entity should linked to the NIL entity, indicating that it is unlinkable.

B Frequency of Annotations in the CGDC Test Data

NE Type	# Non-NIL NEs	# NIL NEs	# Total NEs
Per	17	119	136
Org	62	39	101
Loc	266	64	330
Misc	28	15	43
TOTAL	373	237	610

Table 3: The number of linkable (non-NIL) named entities (NEs) and unlinkable (NIL) NEs in our CGDC test set, broken down by named entity type.

C Performance of EL Models on CGDC Test Subsets

Test set	Model	Non-NAC	NAC	OAC
Morrab	BLINK	0.720	0.326	0.558
	EPGEL	**0.746**	0.615	0.692
	mGENRE	0.631	**0.912**	**0.747**
PCW	BLINK	0.793	0.432	0.656
	EPGEL	**0.817**	0.569	0.725
	mGENRE	0.726	**0.868**	**0.779**

Table 4: EL results on the CGDC test data, broken down by subset. Key: Non-NAC = non-NIL accuracy; NAC = NIL accuracy; OAC = overall accuracy.

D Further Information on Examples with Wrong EL Predictions

	NE with Full Context	URL of Source Item
1	Constance Browne. **Constance Amelia Browne** was the maternal Aunt of Caldwell. She was born on 10th March 1833 at Market Rasen, Lincolnshire, the eldest daughter of Henry Albert...	https://photoarchive.morrablibrary.org.uk/items/show/17268
2	The Bathers/ **Sir James Jebusa Shannon** / 1900-23. Oil painting in the collections of Newport Museum and Art Gallery.	https://www.peoplescollection.wales/items/28049
3	Library. Programme for the Official Opening of Sandfields Branch Library, **Morrison Road**, Sandfields in 1961.	https://www.peoplescollection.wales/items/517688
4	Penzance Theatre Playbill. Advertising the play **Duenna**, to be staged at the New Theatre, Penzance. Location possibly Chapel Street area.	https://photoarchive.morrablibrary.org.uk/items/show/15667
5	VADs, doctors and patients (including many Belgian refugees and Belgian soldiers), outside the old **County Club hotel** (a hospital), now the county library, Beaufort Road, Llandrindod, circa 1915.	https://www.peoplescollection.wales/items/28537

Table 5: Full context for each of the examples shown in Table 2 and the URL of the corresponding CGDC item.