Responsible NLP Checklist

Paper title: *It's All About the Confidence: An Unsupervised Approach for Multilingual Historical Entity Linking using Large Language Models*
Authors: *Cristian Santini, Marieke van Erp, Mehwish Alam*

How to read the checklist symbols:

☑ the authors responded 'yes'

☒ the authors responded 'no'

N/A the authors indicated that the question does not apply to their work

☐ the authors did not respond to the checkbox question

For background on the checklist and guidance provided to the authors, see the Responsible NLP Checklist page at ACL Rolling Review.

☑ A. Questions mandatory for all submissions.

☑ A1. Did you describe the limitations of your work?
This paper has a Limitations section.

☑ A2. Did you discuss any potential risks of your work?
We discuss potential risks in Section 10 (Ethical Considerations). We address the environmental impact of using LLMs for inference, the risk of biases inherent in both the bi-encoder and LLM components (particularly affecting underrepresented languages like Finnish and Swedish), and the potential for propagating errors in knowledge bases like Wikidata that may contain incomplete or culturally biased information about historical entities.

☑ B. Did you use or create scientific artifacts? (e.g. code, datasets, models)

☑ B1. Did you cite the creators of artifacts you used?
The original authors responsible for the creation of the datasets used in our experiments are clearly referenced in Section 4.1 (Evaluation Datasets). The original scupresenting BELA is cited in Section 3 (Methodology).

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We use existing scientific artifacts and make our code publicly available. The source code for MHEL-LLaMo will be made available on GitHub under an MIT license, with the URL reported in Section 1 (Introduction). We use four existing benchmark datasets (HIPE-2020, NewsEye, AJMC, and MHERCL), which are described in Section 4.1 (Evaluation Datasets) along with citations to their original papers and repositories. We also use existing pre-trained models: BELA for candidate retrieval, and instruction-tuned LLMs (Mistral-24B, Gemma-3-27B, and Poro-2-8B) described in Section 3 (Methodology). All model URLs and versions are provided in footnotes throughout the paper to ensure reproducibility.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Our use of existing artifacts is consistent with their intended purposes. The benchmark datasets (HIPE-2020, NewsEye, AJMC, MHERCL) described in Section 4.1 were originally created for

The Responsible NLP Checklist used at ACL Rolling Review is adopted from NAACL 2022, with the addition of ACL 2023 question on AI writing assistance and further refinements based on ARR practice.

evaluating entity linking systems on historical texts, which aligns with our research objectives. The pre-trained models (BELA, Mistral-24B, Gemma-3-27B, Poro-2-8B) mentioned in Section 3 are publicly available instruction-tuned models designed for NLP tasks, consistent with our application. Our source code is released under an MIT license (Section 1) for research purposes. In Section 10 (Ethical Considerations), we discuss potential risks including biases affecting marginalized populations and the limitations of historical entity representation in knowledge bases. We note that our system may privilege well-represented entities in web-crawled corpora, potentially marginalizing references from underrepresented communities.

☑ B4. Did you discuss the steps taken to check whether the data that was collected/used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect/anonymize it?
The datasets used in this study (HIPE-2020, NewsEye, AJMC, MHERCL) are publicly available historical corpora containing texts from the 19th and 20th centuries, which are in the public domain. As described in Section 4.1, these datasets consist of historical press articles, classical commentaries, and music periodicals. The task of entity linking inherently requires the presence of named entities referring to real-world individuals, locations, and organizations, so anonymization would be counterproductive and would render the task meaningless. All individuals mentioned in these historical texts are either historical figures or their references are already in the public domain due to the age of the documents (100+ years old). No modern personal data or living individuals are involved in our study. The datasets were created and curated by their original authors with appropriate considerations for their historical nature, as documented in their respective publications cited in Section 4.1.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
A documentation for our source code is available on Github. An URL is reported in the Section 1, footnote 3 (Introduction).

☑ B6. Did you report relevant statistics like the number of examples, details of train/test/dev splits, etc. for the data that you used/created?
Statistics about the evaluation datasets used are reported in Section 4.1, Table 1 (Evaluation Datasets)

☑ **C. Did you run computational experiments?**

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Model parameters are reported in Section 3 (Methodology), where we specify the use of Mistral-24B (24 billion parameters), Gemma-3-27B (27 billion parameters), and Poro-2-8B (8 billion parameters) as instruction-tuned LLMs. The computational infrastructure and budget are reported in Section 4.2 (Architecture Configuration), where we describe the use of a GPU cluster with two NVIDIA L40S GPUs (246GB) with a total computational budget of approximately 51 GPU hours for running all experiments across the four datasets and six languages.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Experimental Setup is reported in Section 4 (Experimental Setup). Model hyper-parameters are reported in Appendix B.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
To reduce computational costs, we reported results obtained with a single run, this is clearly stated in Section 4.2 (Architecture Configuration).)

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, SpaCy, ROUGE, etc.), did you report the implementation, model, and parameter settings used?
The URL to our source code repository, detailing speicifc package versions used in our experiments, is reported in Section 1 (Introduction).

☒ **D. Did you use human annotators (e.g., crowdworkers) or research with human subjects?**

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
This work re-used previously released datasets, therefore we did not need to use human annotators.

☒ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
This work re-used previously released datasets, therefore we did not need to use human annotators.

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating (e.g., did your instructions explain how the data would be used)?
The re-used datasets are available on public licensed and re-used under the terms of their license.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No ethics review board was included in this research.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No annotators contributed to this research.

☑ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

☑ E1. If you used AI assistants, did you include information about their use?
Yes. AI assistants (e.g., ChatGPT) were used exclusively to improve the readability and clarity of the manuscript through language editing and proofreading. No AI assistance was used for the design, implementation, analysis, or interpretation of the research itself.