## Responsible NLP Checklist

Paper title: *A Domain-Specific Curated Benchmark for Entity and Document-Level Relation Extraction*
Authors: *Marco Martinelli, Stefano Marchesin, Vanessa Bonato, Giorgio Di Nunzio, Nicola Ferro, Ornella Irrera, Laura Menotti, Federica Vezzani, Gianmaria Silvello*

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 limitations and potential risks connected to our work in Sections "Limitations" and "Ethical Considerations".*

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

☑ B1. Did you cite the creators of artifacts you used?
*The models used for pre-annotating documents and as baseline system are cited in Sections 2 (see "First Annotation Phase" and "Automatically-Annotated Data Collection") and 4 (see "Baseline System")*

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
*In "Ethical Considerations" we specify the terms and conditions for use and reuse of the documents included in our collection*

☑ 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)?
*In "Ethical Considerations" we specify the terms and conditions for use and reuse of the documents included in our collection*

☒ 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?
*In "Ethical Considerations" we specify that annotations are released including anonymized identifiers for the annotators, designed to preserve the utility of the data while ensuring that no individual annotator can be personally identified from the released data.*

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?

---

*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.*

*The query employed for document retrieval are presented in Section 2. Moreover, in Tables 6, 7 and 8 we provide detailed information about our annotation schema.*

☑ 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?
*We provide detailed statistics on our annotations in Tables 1,2 and in Figure 4*

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

N/A C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
*(left blank)*

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*Section 4 (see "Baseline System")*

☒ 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?
*Our baseline system uses deterministic approaches, so we dont expect results to vary across runs.*

☑ 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?
*Section 4 (see "Baseline System")*

☑ **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.?
*The annotation guidelines are linked in footnote (4) at page 3*

☑ 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)?
*The recruitment process is discussed in Section 2, while compensation is discussed in Section "Ethical Considerations"*

☑ 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)?
*In Section 2 ("Collection Overview") we discuss the use of anonymized IDs for annotators. No other personal information about annotators has been collected.*

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*(left blank)*

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*In Section 2 ("Collection Overview") we discuss the distinction of annotators in "experts" and "layperson". No other characteristic of the annotator population has been collected.*

☑ **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?
*In Section "Ethical Considerations" we discuss the use of generative AI in our work.*