Responsible NLP Checklist

Paper title: *Coherence of Argumentative Dialogue Snippets: A New Method for Large Scale Evaluation with an Application to Inference Anchoring Theory*
Authors: *Paul Piwek, Jacopo Amidei, Svetlana Stoyanchev*

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.

N/A A2. Did you discuss any potential risks of your work?
This paper reports on an empirical study evaluating a existing theory of dialogue/discourse coherence.

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

☑ B1. Did you cite the creators of artifacts you used?
Moral Maze corpus MM2012c (Section 3.1), Kialo argument maps (Appendix E)

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
Moral Maze corpus MM2012c (Section 3.1) (References to papers through which artefacts are shared), Kialo argument maps (Appendix E)

☑ 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)?
Moral Maze corpus MM2012c (Section 3.1) (References to papers through which artefacts are shared), Kialo argument maps (Appendix E)

N/A 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?
Collected judgements that make up our dataset did not include open text; only coherence rankings on a 7-point scale.

N/A B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We reference relevant papers on the artefacts used.

☑ 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?
Section 3 and Supplementary Materials (Dataset).

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.

☑ **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?
We used a traditional program with symbolic processing with no requirement beyond a standard Windows or Mac machine to run.

N/A C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We provide the algorithms and code (which are traditional, symbolic rather than ML-based).

☑ 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?
Section 4 on Results.

☑ 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?
Sections 3 and 4 - details provided for statistical packages used for data analysis.

☑ **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.?
Appendix F as well as Data Supplementary Materials (Dataset), see file named Qualtrics experiment forms and guidelines.pdf

☑ 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)?
Section 3.2 on Participants.

☑ 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)?
Appendix F as well as Data Supplementary Materials, see file named Qualtrics experiment forms and guidelines.pdf

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Data Data Supplementary Materials, see file named Qualtrics experiment forms and guidelines.pdf See also after Acknowledgements.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section 3.2,

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

N/A E1. If you used AI assistants, did you include information about their use?
(left blank)