Responsible NLP Checklist

Paper title: *Multi-token Mask-filling and Implicit Discourse Relations*
Authors: *Meinan Liu, Yunfang Dong, Xixian Liao, Bonnie Webber*

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 do not discuss potential risks as we are using a dataset (the Wall Street Journal portion of the Penn TreeBank) that has been used in NLP research since 1995, and annotations over that dataset (from the Penn Discourse TreeBank) that has been used in NLP research since 2019.

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

☑ B1. Did you cite the creators of artifacts you used?
Section 7 (data) describes the mapping of Penn TreeBank syntactic parse trees that start with a preposed NP or PP into a parse tree in which the NP or PP has been inserted after the verb. The Penn TreeBank is cited multiple times in the paper.

N/A B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
The terms of use of the Penn TreeBank have been specified by the LDC and have been in effect for ~30 years. The terms of use of the Penn Discourse TreeBank (version 3) are also specified by the LDC. The Penn TreeBank parse trees were originally intended for use in evaluating parsers, in inducing parsers (via machine learning) and in linguistic analysis. The latter is what we have used it for.

☑ 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)?
The Penn TreeBank parse trees were originally intended for use in evaluating parsers, in inducing parsers (via machine learning) and in linguistic analysis. The latter is what we have used it for. Our use of the PTB is described in Section 7. It just involves extracting sentences with preposed NPs or PPs.

☑ 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 Penn TreeBank is built over the Penn WSJ corpus .Individual people are named throughout the

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.

corpus. However, since all the text has been publically available for nearly 30 years, no steps have been taken to anonymize it.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Appendix D of the paper provides examples of sentences whose preposed NP or PP has been moved to canonical position after the verb. Otherwise the data in the artifact is no different from the text in the original WSJ component of the Penn TreeBank.

☑ 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 7 contains statistics on discourse relations implied by mask filling on sentences with Preposed NPs or PPs (called the "Preposed Set"), discourse relations implied by mask filling on those same sentences with the Preposed NP or PP in post-verbal position (called the "Canonical Set"), mask filling on the complement of the Preposed Set, and finally, discourse relations implied by mask filling on just Arg2 of the Preposed Set and just Arg2 of the Canonical Set (called the "Preposed Arg2 Set" and the "Canonical Arg2 Set").

☑ **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?
Section 4 describes the format of the data submitted to BERT under each condition. We used BERT-Base "out of the box", with no change to its parameters in predicting the fillers of a mask inserted between the two arguments of an implicit discourse relation. BERT-Base is noted in Section 4.3 as having 22.9 million trainable parameters. We made no changes.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
The experimental setup is described in Section 4, but none of the experiments included hyperparameter search or best hyperparameter values.

☑ 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 7 ("Results") gives two sets of output statistics for each experiment: Accuracy (what percentage of the top-5 mask fillers can convey a discourse relational sense that matches the "gold label" assigned by human annotators) and "Precision at N" (P@N, what percentage of the top 1, top 2, top 3, top 4 and top 5 mask fillers can convey a discourse relational sense that matches the "gold label" assigned by human annotators).

☒ 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?
We performed very simple calculations for Accuracy and Precision@N. The formulas we used for both are given in the introduction to Section 5.

☒ **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.?
(left blank)

☒ 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)?
(left blank)

☒ N/A 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)?
(left blank)

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

☒ N/A D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
(left blank)

☑ **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?
We used BERT only in our experiments. We did not use any AI assistants in coding or writing.