Responsible NLP Checklist

Paper title: *Effective QA-Driven Annotation of PredicateArgument Relations Across Languages*
Authors: *Jonathan Davidov, Aviv Slobodkin, Shmuel Tomi Klein, Reut Tsarfaty, Ido Dagan, Ayal Klein*

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?
(left blank)

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

☑ B1. Did you cite the creators of artifacts you used?
We acknowledge the creators of all datasets, models, and code used in this work. Specifically: English QA-SRL Parser: We use the T5-based QA-SRL + QANom parser released by Cattan et al. (2024) (see Section 2.2). Universal Dependencies Corpora: Sentences and annotations are drawn from UD treebanks (Nivre et al., 2016; de Marneffe et al., 2021) for Hebrew, Russian, and French (see Section 4). Word Alignment: We employ SimAlign (Jalili Sabet et al., 2020) for bidirectional alignment between source and translated sentences (see Section 3.2). Few-Shot Baselines: GPT-4o and LLaMA-4-Maverick are used as in-context learning baselines (OpenAI, 2024; Meta, 2025) (see Section 5.2). Sentence Similarity Model: A SentenceTransformers paraphrase model is used for semantic question matching (see Appendix H)

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
All datasets, models, and code used in this work are publicly released research artifacts under standard academic or open-source licenses, and we use them unmodified for research purposes only.

☒ 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)?
We did not include a separate discussion of intended use and license compatibility in the paper because all artifacts used and created in this work are permit academic research under standard community datasets and open research model licenses. Their intended use (research and evaluation) is implicit and aligned with prior QA-SRL/QANom and Universal Dependencies work. To avoid redundancy and maintain focus on the methodological contributions, we followed the convention of citing the original resources in the main text and including access details in the supplementary material, without a dedicated license or use-case discussion.

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.

☒ 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?
We did not include a discussion of personally identifiable information (PII) or offensive content filtering because all data used in this work comes from existing Universal Dependencies (UD) corpora. These resources are curated, publicly released datasets that undergo their own anonymization and content review as part of their creation process. Since we did not collect any new data and only used pre-tokenized, non-identifying sentences from UD treebanks, no additional steps for anonymization or PII protection were necessary.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
in section 4.1

☑ 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?
in section 4.1

☑ **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?
in section 5.2 and in Appendix J

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix J

☒ 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?
the test set results reported are a single inference run over the selected model tuned on the dev set

N/A 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?
(left blank)

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

N/A 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.?
We guide the Russian annotator orally on the task

☑ 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)?
in section 4.2

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?**

☒ E1. If you used AI assistants, did you include information about their use?
we use AI assistance for polishing the writing of submitted manuscript, in accordance with ACL policies (e.g. https://2023.aclweb.org/blog/review-acl23/#faq-can-i-use-ai-writing-assistants-to-write-my-review)