Challenges in Reproducing Human Evaluation Results for Role-Oriented Dialogue Summarization

Takumi Ito
Tohoku University
Sendai, Japan;
Utrecht University
the Netherlands
t-ito@tohoku.ac.jp

Qixiang Fang
Utrecht University
the Netherlands
q.fang@uu.nl

Pablo Mosteiro
Utrecht University
the Netherlands
p.mosteiro@uu.nl

Albert Gatt
Utrecht University
the Netherlands
a.gatt@uu.nl

Kees van Deemter
Utrecht University
the Netherlands
c.j.vandeemter@uu.nl

Abstract

There is a growing concern regarding the reproducibility of human evaluation studies in NLP. As part of the ReproHum campaign, we conducted a study to assess the reproducibility of a recent human evaluation study in NLP. Specifically, we attempted to reproduce a human evaluation of a novel approach to enhance Role-Oriented Dialogue Summarization by considering the influence of role interactions. Despite our best efforts to adhere to the reported setup, we were unable to reproduce the statistical results as presented in the original paper. While no contradictory evidence was found, our study raises questions about the validity of the reported statistical significance results, and/or the comprehensiveness with which the original study was reported. In this paper, we provide a comprehensive account of our reproduction study, detailing the methodologies employed, data collection, and analysis procedures. We discuss the implications of our findings for the broader issue of reproducibility in NLP research. Our findings serve as a cautionary reminder of the challenges in conducting reproducible human evaluations and prompt further discussions within the NLP community.

1 Introduction

Natural Language Processing (NLP) has witnessed remarkable advances in recent years. Human evaluation plays a pivotal role in assessing the effectiveness of NLP systems and their performance in meeting specific task requirements. However, concerns have arisen regarding the reproducibility of human evaluation studies in the NLP community (Belz et al., 2022; Huidrom et al., 2022). Reproducibility is defined as the ability of other researchers to repeat the experiments under identical conditions and obtain consistent results.

As part of the ReproHum campaign (Belz and Reiter, 2022), which strives to systematically assess the reproducibility of human evaluation studies in NLP, we conducted a rigorous reproduction study of Lin et al. (2022), with the title *Other Roles Matter! Enhancing Role-Oriented Dialogue Summarization via Role Interactions*.

Dialogue summarization aims to distil relevant information from conversations while preserving their context, presenting a concise and informative summary. The quality of such summarization systems is critical in real-world applications, and a careful evaluation of said quality is a prerequisite to the application of summarization systems.

Lin et al. (2022) start from the idea that, when a system summarises a dialogue between a user (e.g., a customer) and an agent (e.g., someone who answers the customer's questions), it can be helpful to attend to each of these two roles (user, agent) separately. When a user's utterance is summarised, some information from the agent should be taken into account, and the other way around. The authors hypothesise that cross-attention and self-attention can help create an optimal combination of both roles, and they investigate various neural mechanisms for doing so, in particular BERT (Devlin et al., 2019) and PGN (See et al., 2017). After an extensive metric-based evaluation, their human evaluation — on which our paper focuses — tests the hypothesis that, for both BERT and PNG, better summaries are generated when both cross-attention and self-attention are used (in Lin et al. (2022), the systems with these mechanisms are referred to as *BERT-both* and *PNG-both*), compared to settings where the dialogue is summarised as one whole without separating the two roles (settings referred to as *BERT-multi* and *PNG-multi*). They conclude that adding both interactions increases performance with respect to the baseline case.

The objective of our reproduction study was to validate the reported statistical results from the original paper and to investigate the reproducibility of the human evaluation process as outlined by the authors. To achieve this, we meticulously replicated the experimental setup provided in the original work, while also seeking clarification from the authors regarding details of their experimental procedure.

In this paper, we present our findings from the reproduction study, shedding light on the challenges and implications of conducting reproducible human evaluations in the NLP domain.

Our study uncovers significant discrepancies between the statistical results reported in the original paper and those obtained in our reproduction attempt. While we did not find any contradicting evidence, our results raise questions about the validity of the original statistical-significance findings. We emphasise that our aim is not to undermine a valuable piece of work, but to contribute to the ongoing discussion on reproducibility, fostering a more transparent and reliable foundation for future advancements.

2 Data

For our evaluation, we used 100 sample dialogues from the same Chinese Sales Dialogue Summarization (CSDS) dataset used in Lin et al. (2022). The samples were provided to us by the ReproHum organizers. For each of the 100 dialogues, there are two kinds of summaries (user and agent) generated by each of the following four systems: PGN-multi, PGN-both, BERT-multi, BERT-both. Thus, there are 800 summaries in total. A sub-summary refers to a complete sentence in the role-oriented summary.

3 Experimental Setup

We closely follow the guidelines provided in the original paper by Lin et al. (2022). We asked participants to assess the summary quality of Role-Oriented Dialogue Summarization models on three aspects: informativeness, non-redundancy, and fluency. We sought to replicate the evaluation process as faithfully as possible, while also addressing certain details that were obtained from the original authors but were not explicitly mentioned in the original paper.

We treated each "sub-summary" (i.e. sentence) in the role-oriented summary as an individual unit to be scored by the annotators. The evaluation was carried out by three trained volunteers who were familiarised with the evaluation rules provided by the original authors. In the original study, annotators were graduate school-level students and spoke native Mandarin. They were recruited from among the members of the lab conducting the study. In a similar spirit, we recruited three PhD candidates from the department of Information and Computing Sciences at Utrecht University, all of whom self-reported Mandarin as their native language. Contrary to the original study where the annotators were not paid for their participation, we will pay each of our annotators 120 Euros for 12 hours of work.[1]

The annotators assessed each sub-summary according to three pre-defined aspects: informativeness, non-redundancy, and fluency. Each sub-summary received a score for each aspect based on the perceived quality of the summary with respect to that particular aspect.

As was done in the original paper, we first gave the three annotators the same ten summaries, and asked them to rate those summaries. To ensure the reliability of the obtained scores, we conducted an inter-annotator agreement analysis. This process involved comparing the scores given by the three volunteers for each sub-summary. We used Cohen's kappa coefficient as a measure of agreement. This was calculated by concatenating all values for each participant together.

We then gave each participant a different set of 30 summaries to rate. In total, there were 100 summaries: 10 were annotated by all three participants, while the remaining 90 were annotated by one participant each.

To represent the summary quality in general, we aggregated the scores for all three aspects (informativeness, non-redundancy, and fluency) into an "Overall" metric for each sub-summary. The overall score for a sub-summary was obtained by averaging the individual scores assigned by the annotators for that specific aspect.

The obtained scores were then normalised to a range between 0 and 1 to facilitate comparison and presentation. The normalised scores were compiled into a table, which is analogous to Table 4 in

[1] Payment is still being processed at the time of writing this article.

the original paper, showcasing the performance of different models across the evaluated aspects.

There is some ambiguity regarding how scores should be computed for the summaries that were evaluated by the three participants. In particular, every sub-summary evaluated by a single participant has a single score; but for the summaries evaluated by all participants (which was done for the purpose of computing the inter-annotator agreement), there are three scores per sub-summary. The original paper does not specify how the scores were computed for these *multi-annotated* summaries. We performed our analysis under four "cases". These are defined by the way we compute the score for each multi-annotated summary:

1. use the scores of participant 1
2. use the scores of participant 2
3. use the scores of participant 3
4. use the average score among participants

Although we felt that it was necessary to distinguish between these four cases, we will see in Section 4 that our overall conclusions do not depend on which case we focus on.

To ensure transparency and to facilitate reproducibility of our study, we have made our code and datasets publicly available on our GitHub repository[2]. The repository contains the necessary scripts and documentation to replicate our experimental procedures and results accurately.

4 Results and Discussion

4.1 Participant results

After the first 10 annotations, the results of the three annotators were compared, and we calculated the inter-annotator agreement using Cohen's kappa, as in the original paper. We computed κ for each pair of annotators, and computed the average of the three values. We obtained a κ_{average}=0.48. This was exactly the same – admittedly rather low (see Section 5 for a discussion) – value as the one reported in the original paper. We then gave 30 more summaries to each annotator, which resulted in a total of 100 summaries being evaluated.

The participants' results presented by Lin et al. (2022) are found in Table 1. These are found in Table 4 of the original paper, and copied here without modification. Table 1 also presents the results of our reproduction experiment. Each horizontal block represents a different case of whose values should be taken for the first 10 annotations; these are referred to as "cases" in Section 3.

4.2 Reproducibility assessment

To assess the reproducibility of the original result, we computed three scores:

1. The Pearson correlation coefficient
2. The fraction of matching both/multi pairs
3. The F1 score of statistical significance results

Further details about each of these follow.

Pearson correlation coefficient. If the results of our experiment reproduced the original experiment exactly, we would have a perfectly linear correlation between the two sets of results. To estimate how far we are from that, we have concatenated all the values in each of the 5 tables of results (the original paper, plus our 4 "cases"), from left to right and top to bottom, and computed the Pearson correlation coefficient between each of our 4 cases and the original paper. The results are shown on Table 2.

Fraction of matching both/multi pairs. The original paper reports a number in boldface if it is larger than its multi/both counterpart. In other words, it highlights the performance of multi vs both, or vice versa. Thus, we have computed, for each of the four cases, the fraction of multi-/both pairs that follow the same trend (lower/equal/higher) as in the original paper. We call this the *matching accuracy A*. It is reported on Table 2.

F1 score of statistical significance. The aforementioned matching accuracy penalises non-matches too harshly, because it does not account for near-matches. Indeed, we are often only interested in the difference between two values if they are statistically significant. To that end, we have computed the F1 score for statistical significance. We consider the original paper as the gold standard. For each value, we take the true label to be 1 if the value is statistically significantly larger than its multi/both counterpart, and 0 otherwise. The results are reported on Table 2. While there exists a reasonable degree of concordance between the numerical values in the original findings and our

[2] https://github.com/taku-ito/reprohum-utrecht

CSDS	Info	Non-Red	Flu	Overall
Lin et al. (2022)				
PGN-multi	**0.69**/0.65	0.54/0.55	0.70/0.79	0.64/0.66
PGN-both	0.66/**0.69**	**0.58/0.59***	**0.73/0.81**	**0.66/0.70***
BERT-multi	0.58/0.56	**0.66/0.61**	0.84/**0.87**	0.69/0.68
BERT-both	**0.62*/0.60***	0.62/0.60	**0.85/0.87**	**0.70/0.69**
Case 1				
PGN-multi	**0.63**/0.59	0.58/0.55	**0.69**/0.70	0.63/0.61
PGN-both	0.62/**0.64***	**0.61/0.59**	0.68/**0.74**	**0.64/0.65***
BERT-multi	0.55/0.45	**0.69*/0.61**	**0.82/0.80**	**0.69*/0.62**
BERT-both	**0.56/0.47**	0.62/0.58	0.78/**0.80**	0.65/**0.62**
Case 2				
PGN-multi	**0.62**/0.58	0.57/0.56	**0.68**/0.69	0.62/0.61
PGN-both	0.61/**0.62**	**0.60/0.58**	0.67/**0.71**	**0.63/0.64***
BERT-multi	**0.55**/0.45	**0.70*/0.60**	**0.82/0.78**	**0.69*/0.61**
BERT-both	**0.55/0.47**	0.62/0.57	0.78/**0.78**	0.65/**0.61**
Case 3				
PGN-multi	**0.64**/0.60	0.59/0.58	**0.69**/0.72	**0.64**/0.63
PGN-both	0.63/**0.65***	**0.62/0.60**	0.68/**0.75**	**0.64/0.67***
BERT-multi	**0.57**/0.46	**0.72*/0.62**	**0.83/0.81**	**0.71*/0.63**
BERT-both	**0.57/0.49**	0.63/0.59	0.79/0.80	0.67/**0.63**
Case 4				
PGN-multi	**0.63**/0.59	0.58/0.56	**0.69**/0.70	0.63/0.62
PGN-both	0.62/**0.64***	**0.61/0.59**	0.68/**0.73**	**0.64/0.65***
BERT-multi	**0.56**/0.45	**0.71*/0.61**	**0.82/0.80**	**0.70*/0.62**
BERT-both	**0.56/0.48**	0.62/0.58	0.78/0.79	0.66/**0.62**

Table 1: Results of Lin et al. (2022), reproduced here without modification (above the double line), along with the results of the present human evaluation (below the double line) under the four “cases” (see Section 3). Each cell contains two numbers separated by a slash: the left number corresponds to the user, and the right number corresponds to the agent. A number for “multi” in boldface indicates that the performance is better than the corresponding number for “both”, and vice versa; if both are the same, both appear in boldface. An asterisk indicates that the difference between the “both” and “multi” results is statistically significant.

Case	Pearson's r	A	F1
1	0.90	0.75	0.25
2	0.89	0.69	0.29
3	0.90	0.56	0.25
4	0.90	0.62	0.25

Table 2: Reproducibility scores between the results of the original experiment and our results. The "cases" refer to how the scores of the ten summaries that were rated by all three participants were aggregated. Pearson's r is computed across all 32 reported values. A: matching accuracy, the fraction of multi/both pairs that follow the same trend (lower/equal/higher) as in the original paper. F1 score is computed by taking the paper as gold standard, labelling a value as 1 if it is statistically significantly larger than its multi/both counterpart, 0 otherwise.

outcomes, as shown by the aforementioned r and A metrics, a notably weaker concurrence is evident when considering the statistical-significance F1 score. This indicates potential issues concerning the efficacy of the employed statistical significance testing methodology. Further elaboration on this matter will be provided in Section 5. We note that, despite the low agreement in the statistical significance of the results, none of the multi/both pairs deemed to be statistically significantly different in the original paper exhibited the opposite trend in our study.

4.3 Comparison of findings

In the original paper, the authors conclude from the human evaluations that applying interactions on the PGN architecture (i.e., using the "both" model) leads to improvements in all metrics except informativeness, where they deem the two options comparable. Meanwhile, for the BERT architecture, the "both" model is better on all metrics except non-redundancy, for which "multi" is better. They also conclude that, given that the "Overall" metric is higher for "both" in both architectures (PGN and BERT), the "both" option is better than "multi".

In our study, the most salient differences are:

- For Fluency+User, PGN-both was worse than PGN-multi in all four cases;
- For Fluency, BERT-both was worse than or equal to BERT-multi for both roles in all four cases;
- For Overall, BERT-both was worse than or equal to BERT-multi for both roles in all four cases

These differences suggest that we cannot reproduce the original paper's conclusion that "both" is generally better than "multi", at least based on the human evaluation alone.

5 Conclusions

In this paper, we conducted a reproduction study of the human evaluation in Lin et al. (2022), as part of the ReproHum campaign to assess the reproducibility of human evaluation in NLP (Belz et al., 2023). Our objective was to assess the reproducibility of the results reported in the original paper and thoroughly investigate the difference between our results and the original paper's, if any.

Throughout our study, we sought to adhere closely to the original experimental setup. However, our findings reveal notable discrepancies in the statistical results obtained, particularly in comparing the improvements of the "both" method with respect to the "multi" method. In the original paper, "multi" is a baseline method, while "both" adds cross-attention and self-attention interactions to the models (see Section 1). Unlike the original work, our experiments did not demonstrate clear improvements in summary quality when considering role interactions.

Despite the differences in our obtained results, we acknowledge the high Pearson correlation coefficient between the original paper's scores and our own, indicating a strong consistency in the relative ranking of models across the evaluation aspects. Furthermore, while our findings were different from the original results in terms of statistical significance, we acknowledge that they are not contradictory, i.e., there is no model for which the authors of the original paper claim statistically significantly better results for "both" or "multi", while we find the opposite to be true (i.e., statistically significantly worse results). We believe that the statistical significance analysis employed in the original paper may have certain flaws. Firstly, we maintain that a correction procedure for inflated type-1 error should have been applied, considering that multiple statistical significance tests were conducted on the same dataset. Failure to account for this potential bias might have resulted in too many false positives (i.e. results which appear to be statistically significant but are not). Secondly, the authors computed statistical significance tests, but then also drew conclusions from results that were not statistically significantly different. This should be avoided.

One significant observation from our study was the relatively low level of agreement among the annotators. This raises concerns about the consistency of the evaluation process and the potential for different interpretations of the instructions. It would have been valuable to closely scrutinise the reasons for such disagreement. If the disagreements stemmed from differing interpretations of the guidelines, an update to the instructions and a restart of the annotation process might have been necessary. Alternatively, if the disagreements were legitimate, the study could have been improved by having multiple annotators assess all the summaries, allowing for a better understanding of the inherent variability. This is along the lines of recent work that tries to account for inherent variability when training NLP models (Leonardelli et al., 2023).

Regarding the reproducibility experiment itself, the description provided in the original paper was insufficient for us to fully attempt a replication. Nonetheless, thanks to the cooperation of the authors, we were able to clarify the necessary procedures. Even so, we had to perform four studies under different "cases", which refer to the various ways we pooled together the results of the first 10 summaries, annotated by all participants.

Moreover, the data collection process posed significant challenges, largely due to the participants making multiple errors that needed to be corrected before statistical analysis became feasible. Namely, we observed several mismatches between the number of sentences and the number of annotations provided by participants. This was probably caused by the annotation being done in a spreadsheet. In those cases, we had to ask participants to correct their work. Ensuring data quality and accuracy is crucial in human evaluation studies, and these difficulties further underscore the importance of transparent reporting and careful handling of data.

Finally, we wish to clarify that our focus was on the parts of the original paper that dealt with human evaluations, particularly in terms of reproducibility. We do not make any general claims about the strength of the entire original paper, which included metric-based evaluations as well. The results of the metric-based evaluation in the original work may indeed be more convincing.

In conclusion, our reproduction study highlights the importance of carefully reporting the conditions under which a human evaluation was conducted to enhance reproducibility, and the need for thorough reporting of experimental details necessary for reproduction studies, as well as scrutiny of statistical significance analyses in NLP research. We also provide suggestions for future studies to enhance the reproducibility and transparency of human evaluation experiments. Despite the challenges we encountered, we commend the authors for their cooperation, which allowed us to perform a comprehensive reproduction of their work. We believe that open dialogue and collaborative efforts within the research community are essential for advancing the field of NLP and achieving meaningful progress in dialogue summarization and other language generation tasks.

References

Anya Belz, Maja Popovic, and Simon Mille. 2022. Quantified reproducibility assessment of NLP results. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16–28, Dublin, Ireland. Association for Computational Linguistics.

Anya Belz and Ehud Reiter. 2022. ReproHum: Investigating Reproducibility of Human Evaluations in Natural Language Processing. https://gow.epsrc.ukri.org/NGBOViewGrant.aspx?GrantRef=EP/V05645X/1.

Anya Belz, Craig Thomson, and Ehud Reiter. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rudali Huidrom, Ondřej Dušek, Zdeněk Kasner, Thiago Castro Ferreira, and Anya Belz. 2022. Two reproductions of a human-assessed comparative evaluation of a semantic error detection system. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 52–61, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Elisa Leonardelli, Alexandra Uma, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, and Massimo Poesio. 2023. SemEval-2023 Task 11: Learning With Disagreements (LeWiDi).

Haitao Lin, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2022. Other Roles Matter! Enhancing Role-Oriented Dialogue Summarization via Role Interactions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2545–2558, Dublin, Ireland. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

A HEDS sheet

```
{
   "heds-criteria-criterion-response_elicitation-scale_presented_as-5": {
      "data": {
         "Informativeness": true,
         "Fluency": true,
         "Non-redundancy": true
      },
      "control": {},
      "text": {
         "Informativeness": "5. Other (please describe)",
         "Fluency": "5. Other (please describe)",
         "Non-redundancy": "5. Other (please describe)"
      }
   },
   "heds-criteria-criterion-response_elicitation-response_aggregation": {
      "data": {
         "Informativeness": "average the set of per-sentence values.",
         "Non-redundancy": "average the set of per-sentence values.",
         "Fluency": "average the set of per-sentence values."
      },
      "control": {}
   },
   "heds-criteria-criterion-evaluation_mode-objective_or_subjective-1": {
      "data": {
         "Informativeness": true,
         "Non-redundancy": true,
         "Fluency": true
      },
      "control": {},
      "text": {
         "Informativeness": "1. Objective",
         "Non-redundancy": "1. Objective",
         "Fluency": "1. Objective"
      }
   },
   "heds-paper_and_resources-names_and_affiliations-person_completing_this_sheet-affiliation": {
      "data": {
         "": "Utrecht University / Tohoku University"
      },
      "control": {}
   },
   "heds-paper_and_resources-names_and_affiliations-person_completing_this_sheet-name": {
      "data": {
         "": "Takumi Ito"
      },
      "control": {}
   },
   "heds-criteria-criterion-criteria-output_aspect-1": {
      "data": {
         "Informativeness": false,
         "Non-redundancy": false,
         "Fluency": true
      },
      "control": {},
      "text": {
         "Informativeness": "",
         "Non-redundancy": "",
         "Fluency": "1. Form of output"
```

```
    }
},
"heds-sample_evaluators_design-evaluators-evaluators-expertise-other_text": {
    "data": {
        "": ""
    },
    "control": {}
},
"heds-sample_evaluators_design-evaluators-evaluators-expertise-1": {
    "data": {
        "": false
    },
    "control": {},
    "text": {
        "": ""
    }
},
"heds-criteria-criterion-response_elicitation-form_of_response-10": {
    "data": {
        "Informativeness": false,
        "Non-redundancy": false,
        "Fluency": false
    },
    "control": {},
    "text": {
        "Informativeness": "",
        "Non-redundancy": "",
        "Fluency": ""
    }
},
"heds-criteria-criterion-response_elicitation-scale_presented_as-2": {
    "data": {
        "Informativeness": false,
        "Fluency": false,
        "Non-redundancy": false
    },
    "control": {},
    "text": {
        "Informativeness": "",
        "Fluency": "",
        "Non-redundancy": ""
    }
},
"heds-system-input_types-8": {
    "data": {
        "": true
    },
    "control": {},
    "text": {
        "": "8. text: dialogue"
    }
},
"heds-paper_and_resources-names_and_affiliations-contact_author-affiliation": {
    "data": {
        "": "Utrecht University"
    },
    "control": {}
},
"heds-criteria-criterion-response_elicitation-form_of_response-7": {
    "data": {
      "Informativeness": false,
      "Non-redundancy": false,
      "Fluency": false
    },
    "control": {},
    "text": {
      "Informativeness": "",
      "Non-redundancy": "",
      "Fluency": ""
    }
  },
  "heds-sample_evaluators_design-experimental_design-experimental_conditions-8": {
    "data": {
      "": false
    },
    "control": {},
    "text": {
      "": ""
    }
  },
  "heds-system-input_languages-29": {
    "data": {
      "": true
    },
    "control": {},
    "text": {
      "": "29. Chinese"
    }
  },
  "heds-criteria-criterion-response_elicitation-form_of_response-6": {
    "data": {
      "Informativeness": false,
      "Non-redundancy": false,
      "Fluency": false
    },
    "control": {},
    "text": {
      "Informativeness": "",
      "Non-redundancy": "",
      "Fluency": ""
    }
  },
  "heds-sample_evaluators_design-evaluators-evaluators-payment-3": {
    "data": {
      "": false
    },
    "control": {},
    "text": {
      "": ""
    }
  },
  "heds-criteria-criterion-response_elicitation-form_of_response-5": {
    "data": {
      "Informativeness": false,
      "Non-redundancy": false,
      "Fluency": false
    },
    "control": {},
    "text": {
      "Informativeness": "",
      "Non-redundancy": "",
      "Fluency": ""
    }
  },
  "heds-sample_evaluators_design-sample-system_output_selection-1": {
    "data": {
      "": false
    },
    "control": {},
    "text": {
      "": ""
    }
  },
  "heds-sample_evaluators_design-experimental_design-quality_assurance-description": {
    "data": {
      "": "N/A"
    },
    "control": {}
  },
  "heds-sample_evaluators_design-evaluators-evaluators-expertise-2": {
    "data": {
      "": true
    },
    "control": {},
    "text": {
      "": "2. non-experts"
    }
  },
  "heds-sample_evaluators_design-experimental_design-experimental_conditions-1": {
    "data": {
      "": true
    },
    "control": {},
    "text": {
      "": "1. evaluation carried out by evaluators at a place of their own choosing, e.g. online, using a paper form, etc."
    }
  },
  "heds-ethics-review_body": {
    "data": {
      "": "No"
    },
    "control": {}
  },
  "heds-criteria-criterion-response_elicitation-effect_size_method": {
    "data": {
      "Non-redundancy": ""
    },
    "control": {}
  },
  "heds-criteria-criterion-response_elicitation-verbatim_question": {
    "data": {
      "Informativeness": "N/A ",
      "Non-redundancy": "N/A",
      "Fluency": "N/A"
    },
    "control": {}
  },
  "heds-sample_evaluators_design-evaluators-evaluators-known_to_authors-2": {
    "data": {
        "": false
    },
    "control": {},
    "text": {
        "": ""
    }
},
"heds-criteria-criterion-evaluation_mode-intrinsic_or_extrinsic-1": {
    "data": {
        "Informativeness": false,
        "Non-redundancy": true,
        "Fluency": true
    },
    "control": {},
    "text": {
        "Informativeness": "",
        "Non-redundancy": "1. Intrinsic",
        "Fluency": "1. Intrinsic"
    }
},
"heds-criteria-criterion-evaluation_mode-objective_or_subjective-other_text": {
    "data": {
        "Informativeness": "",
        "Non-redundancy": "",
        "Fluency": ""
    },
    "control": {}
},
"heds-criteria-criterion-response_elicitation-scale_presented_as-other_text": {
    "data": {
        "Informativeness": "fill in the cells on the spreadsheet",
        "Non-redundancy": "fill in the cells on the spreadsheet",
        "Fluency": "fill in the cells on the spreadsheet"
    },
    "control": {}
},
"heds-sample_evaluators_design-evaluators-evaluators-payment-1": {
    "data": {
        "": true
    },
    "control": {},
    "text": {
        "": "1. paid (monetary compensation)"
    }
},
"heds-sample_evaluators_design-evaluators-evaluators-description": {
    "data": {
        "": "Chinese PhD candidates in the same department as the authors."
    },
    "control": {}
},
"heds-sample_evaluators_design-sample-number_of_system_outputs": {
    "data": {
        "": "100"
    },
    "control": {}
},
"heds-system-output_types-4": {
    "data": {
        "": false
    },
    "control": {},
    "text": {
        "": ""
    }
},
"heds-criteria-criterion-evaluation_mode-objective_or_subjective-2": {
    "data": {
        "Informativeness": false,
        "Non-redundancy": false,
        "Fluency": false
    },
    "control": {},
    "text": {
        "Informativeness": "",
        "Non-redundancy": "",
        "Fluency": ""
    }
},
"heds-criteria-criterion-response_elicitation-form_of_response-4": {
    "data": {
        "Informativeness": false,
        "Non-redundancy": false,
        "Fluency": false
    },
    "control": {},
    "text": {
        "Informativeness": "",
        "Non-redundancy": "",
        "Fluency": ""
    }
},
"heds-criteria-criterion-response_elicitation-size_of_scale-2": {
    "data": {
        "Informativeness": false,
        "Non-redundancy": false,
        "Fluency": false
    },
    "control": {},
    "text": {
        "Informativeness": "",
        "Non-redundancy": "",
        "Fluency": ""
    }
},
"heds-sample_evaluators_design-evaluators-evaluators-known_to_authors-other_text": {
    "data": {
        "": ""
    },
    "control": {}
},
"heds-criteria-criterion-criteria-output_aspect-other_text": {
    "data": {
        "Informativeness": "",
        "Non-redundancy": "",
        "Fluency": ""
    },
    "control": {}
},
"heds-criteria-criterion-evaluation_mode-absolute_or_relative-1": {
  "data": {
    "Informativeness": false,
    "Non-redundancy": false,
    "Fluency": false
  },
  "control": {},
  "text": {
    "Informativeness": "",
    "Non-redundancy": "",
    "Fluency": ""
  }
},
"heds-sample_evaluators_design-experimental_design-quality_assurance-method-1": {
  "data": {
    "": true
  },
  "control": {},
  "text": {
    "": "1. evaluators are required to be native speakers of the language they evaluate."
  }
},
"heds-sample_evaluators_design-evaluators-evaluators-payment-other_text": {
  "data": {
    "": ""
  },
  "control": {}
},
"heds-sample_evaluators_design-experimental_design-experimental_conditions-other_text": {
  "data": {
    "": ""
  },
  "control": {}
},
"heds-criteria-criterion-criteria-quality_type-2": {
  "data": {
    "Informativeness": false,
    "Non-redundancy": false,
    "Fluency": true
  },
  "control": {},
  "text": {
    "Informativeness": "",
    "Non-redundancy": "",
    "Fluency": "2. Goodness"
  }
},
"heds-criteria-criterion-response_elicitation-form_of_response-9": {
  "data": {
    "Informativeness": false,
    "Non-redundancy": false,
    "Fluency": false
  },
  "control": {},
  "text": {
    "Informativeness": "",
    "Non-redundancy": "",
    "Fluency": ""
  }
},
"heds-criteria-criterion-evaluation_mode-intrinsic_or_extrinsic-2": {
      "data": {
        "Informativeness": true,
        "Non-redundancy": false,
        "Fluency": false
      },
      "control": {},
      "text": {
        "Informativeness": "2. Extrinsic",
        "Non-redundancy": "",
        "Fluency": ""
      }
    },
    "heds-sample_evaluators_design-sample-system_output_selection-other_text": {
      "data": {
        "": "The same samples used in the original paper."
      },
      "control": {}
    },
    "heds-sample_evaluators_design-sample-system_output_selection-5": {
      "data": {
        "": true
      },
      "control": {},
      "text": {
        "": "5. other (please describe)"
      }
    },
    "heds-sample_evaluators_design-experimental_design-evaluator_freedom-other_text": {
      "data": {
        "": "No restrictions."
      },
      "control": {}
    },
    "heds-criteria-criterion-criteria-quality_type-1": {
      "data": {
        "Informativeness": true,
        "Non-redundancy": true,
        "Fluency": false
      },
      "control": {},
      "text": {
        "Informativeness": "1. Correctness",
        "Non-redundancy": "1. Correctness",
        "Fluency": ""
      }
    },
    "heds-criteria-criterion-evaluation_mode-absolute_or_relative-2": {
      "data": {
        "Informativeness": true,
        "Non-redundancy": true,
        "Fluency": true
      },
      "control": {},
      "text": {
        "Informativeness": "2. Relative",
        "Non-redundancy": "2. Relative",
        "Fluency": "2. Relative"
      }
    },
    "heds-paper_and_resources-names_and_affiliations-contact_author-name": {
      "data": {
        "": "Kees van Deemter"
      },
      "control": {}
    },
    "heds-system-output_types-8": {
      "data": {
        "": false
      },
      "control": {},
      "text": {
        "": ""
      }
    },
    "heds-ethics-special_category_data": {
      "data": {
        "": "No"
      },
      "control": {}
    },
    "heds-sample_evaluators_design-experimental_design-experimental_conditions-6": {
      "data": {
        "": false
      },
      "control": {},
      "text": {
        "": ""
      }
    },
    "heds-system-output_languages-29": {
      "data": {
        "": true
      },
      "control": {},
      "text": {
        "": "29. Chinese"
      }
    },
    "heds-criteria-criterion-response_elicitation-list_or_range": {
      "data": {
        "Informativeness": "0,1,2",
        "Non-redundancy": "0,1,2",
        "Fluency": "0,1,2"
      },
      "control": {}
    },
    "heds-criteria-criterion-response_elicitation-task_description": {
      "data": {
        "Informativeness": "N/A",
        "Non-redundancy": "N/A",
        "Fluency": "N/A"
      },
      "control": {}
    },
    "heds-criteria-criterion-response_elicitation-participant_criterion_name": {
      "data": {
        "Informativeness": "Informativeness",
        "Non-redundancy": "Non-redundancy",
        "Fluency": "Flunecy"
      },
    "control": {}
},
"heds-sample_evaluators_design-evaluators-recruitment_method": {
    "data": {
        "": "sent an email to those who met the requirements"
    },
    "control": {}
},
"heds-paper_and_resources-names_and_affiliations-contact_author-email": {
    "data": {
        "": "c.j.vandeemter@uu.nl"
    },
    "control": {}
},
"heds-sample_evaluators_design-evaluators-evaluators-known_to_authors-1": {
    "data": {
        "": true
    },
    "control": {},
    "text": {
        "": "1. previously known to authors"
    }
},
"heds-criteria-criterion-response_elicitation-form_of_response-2": {
    "data": {
        "Informativeness": true,
        "Non-redundancy": true,
        "Fluency": true
    },
    "control": {},
    "text": {
        "Informativeness": "2. direct quality estimation",
        "Non-redundancy": "2. direct quality estimation",
        "Fluency": "2. direct quality estimation"
    }
},
"heds-sample_evaluators_design-experimental_design-experimental_conditions-7": {
    "data": {
        "": false
    },
    "control": {},
    "text": {
        "": ""
    }
},
"heds-sample_evaluators_design-evaluators-evaluators-are_authors-2": {
    "data": {
        "": true
    },
    "control": {},
    "text": {
        "": "2. evaluators do not include any of the authors"
    }
},
"heds-sample_evaluators_design-experimental_design-evaluators_place_of_choosing": {
    "data": {
        "": "N/A"
    },
    "control": {}
},
    "heds-sample_evaluators_design-sample-system_output_selection-3": {
        "data": {
            "": false
        },
        "control": {},
        "text": {
            "": ""
        }
    },
    "heds-criteria-criterion": {
        "data": {},
        "control": {
            "Informativeness": false,
            "Non-redundancy": false,
            "Fluency": true
        },
        "text": {}
    },
    "heds-sample_evaluators_design-evaluators-training_practice": {
        "data": {
            "": "ask the participants to read the task description provided by the original authors before starting
the annotation."
        },
        "control": {}
    },
    "heds-criteria-criterion-criteria-quality_type-other_text": {
        "data": {
            "Informativeness": "",
            "Non-redundancy": "",
            "Fluency": ""
        },
        "control": {}
    },
    "heds-sample_evaluators_design-evaluators-evaluators-payment-4": {
        "data": {
            "": false
        },
        "control": {},
        "text": {
            "": ""
        }
    },
    "heds-criteria-criterion-criteria-quality_type-3": {
        "data": {
            "Informativeness": false,
            "Non-redundancy": false,
            "Fluency": false
        },
        "control": {},
        "text": {
            "Informativeness": "",
            "Non-redundancy": "",
            "Fluency": ""
        }
    },
    "heds-criteria-criterion-criteria-self_vs_external_frame-3": {
        "data": {
            "Informativeness": false,
            "Non-redundancy": false,
            "Fluency": false
        },
        "control": {},
        "text": {
            "Informativeness": "",
            "Non-redundancy": "",
            "Fluency": ""
        }
    },
    "heds-sample_evaluators_design-experimental_design-evaluator_freedom-2": {
        "data": {
            "": false
        },
        "control": {},
        "text": {
            "": ""
        }
    },
    "heds-sample_evaluators_design-evaluators-evaluators-are_authors-1": {
        "data": {
            "": false
        },
        "control": {},
        "text": {
            "": ""
        }
    },
    "heds-paper_and_resources-paper-link": {
        "data": {
            "": "https://aclanthology.org/2022.acl-long.182/"
        },
        "control": {}
    },
    "heds-criteria-criterion-evaluation_mode-intrinsic_or_extrinsic-other_text": {
        "data": {
            "Informativeness": "",
            "Non-redundancy": "",
            "Fluency": ""
        },
        "control": {}
    },
    "heds-ethics-personal_data": {
        "data": {
            "": "No"
        },
        "control": {}
    },
    "heds-sample_evaluators_design-experimental_design-experimental_conditions-4": {
        "data": {
            "": false
        },
        "control": {},
        "text": {
            "": ""
        }
    },
    "heds-sample_evaluators_design-sample-system_output_selection-2": {
        "data": {
            "": false
        },
        "control": {},
        "text": {
            "": ""
        }
    },
    "heds-criteria-criterion-response_elicitation-form_of_response-8": {
        "data": {
            "Informativeness": false,
            "Non-redundancy": false,
            "Fluency": false
        },
        "control": {},
        "text": {
            "Informativeness": "",
            "Non-redundancy": "",
            "Fluency": ""
        }
    },
    "heds-criteria-criterion-response_elicitation-size_of_scale-3": {
        "data": {
            "Informativeness": false,
            "Non-redundancy": false,
            "Fluency": false
        },
        "control": {},
        "text": {
            "Informativeness": "",
            "Non-redundancy": "",
            "Fluency": ""
        }
    },
    "heds-criteria-criterion-response_elicitation-scale_presented_as-3": {
        "data": {
            "Informativeness": false,
            "Fluency": false,
            "Non-redundancy": false
        },
        "control": {},
        "text": {
            "Informativeness": "",
            "Fluency": "",
            "Non-redundancy": ""
        }
    },
    "heds-sample_evaluators_design-experimental_design-collection_method": {
        "data": {
            "": "Excel spreadsheet"
        },
        "control": {}
    },
    "heds-criteria-criterion-response_elicitation-inter_annotator-agreement-other_text": {
        "data": {
            "Informativeness": "Cohen's kappa",
            "Non-redundancy": "Cohen's kappa",
            "Fluency": "Cohen's kappa"
        },
        "control": {}
    },
    "heds-system-tasks-16": {
        "data": {
            "": true
    },
    "control": {},
    "text": {
        "": "16. summarisation (text-to-text)"
    }
},
"heds-criteria-criterion-criteria-output_aspect-2": {
    "data": {
        "Informativeness": true,
        "Non-redundancy": false,
        "Fluency": false
    },
    "control": {},
    "text": {
        "Informativeness": "2. Content of output",
        "Non-redundancy": "",
        "Fluency": ""
    }
},
"heds-sample_evaluators_design-evaluators-number_of_evaluators": {
    "data": {
        "": "3"
    },
    "control": {}
},
"heds-sample_evaluators_design-experimental_design-preregistered-2": {
    "data": {
        "": true
    },
    "control": {},
    "text": {
        "": "2. no"
    }
},
"heds-criteria-criterion-criteria-self_vs_external_frame-1": {
    "data": {
        "Informativeness": false,
        "Non-redundancy": false,
        "Fluency": true
    },
    "control": {},
    "text": {
        "Informativeness": "",
        "Non-redundancy": "",
        "Fluency": "1. Quality of output in its own right"
    }
},
"heds-sample_evaluators_design-evaluators-evaluators-payment-2": {
    "data": {
        "": false
    },
    "control": {},
    "text": {
        "": ""
    }
},
"heds-sample_evaluators_design-experimental_design-preregistered-1": {
    "data": {
        "": false
    },
        "control": {},
        "text": {
            "": ""
        }
    },
    "heds-system-output_types-5": {
        "data": {
            "": true
        },
        "control": {},
        "text": {
            "": "5. text: sentence"
        }
    },
    "heds-sample_evaluators_design-experimental_design-preregistered-other_text": {
        "data": {
            "": ""
        },
        "control": {}
    },
    "heds-paper_and_resources-resources-links": {
        "data": {
            "": "https://drive.google.com/drive/u/0/folders/1hevFqMAwx9qZpfvsYSar6e4IBgFuSVKw"
        },
        "control": {}
    },
    "heds-criteria-criterion-criteria-self_vs_external_frame-other_text": {
        "data": {
            "Informativeness": "",
            "Non-redundancy": "",
            "Fluency": ""
        },
        "control": {}
    },
    "heds-sample_evaluators_design-evaluators-evaluators-expertise-3": {
        "data": {
            "": false
        },
        "control": {},
        "text": {
            "": ""
        }
    },
    "heds-sample_evaluators_design-sample-statistical_power-value": {
        "data": {
            "": "N/A"
        },
        "control": {}
    },
    "heds-criteria-criterion-response_elicitation-inter_annotator-agreement-3": {
        "data": {
            "Informativeness": false,
            "Non-redundancy": false,
            "Fluency": false
        },
        "control": {},
        "text": {
            "Informativeness": "",
            "Non-redundancy": "",
            "Fluency": ""
    }
},
"heds-sample_evaluators_design-evaluators-evaluators-are_authors-other_text": {
    "data": {
        "": ""
    },
    "control": {}
},
"heds-sample_evaluators_design-experimental_design-experimental_conditions-3": {
    "data": {
        "": false
    },
    "control": {},
    "text": {
        "": ""
    }
},
"heds-criteria-criterion-response_elicitation-form_of_response-1": {
    "data": {
        "Informativeness": false,
        "Non-redundancy": false,
        "Fluency": false
    },
    "control": {},
    "text": {
        "Informativeness": "",
        "Non-redundancy": "",
        "Fluency": ""
    }
},
"heds-criteria-criterion-response_elicitation-size_of_scale-other_text": {
    "data": {
        "Informativeness": "3",
        "Non-redundancy": "3",
        "Fluency": "3"
    },
    "control": {}
},
"heds-sample_evaluators_design-experimental_design-evaluator_freedom-3": {
    "data": {
        "": true
    },
    "control": {},
    "text": {
        "": "3. neither of the above (please describe)"
    }
},
"heds-criteria-criterion-response_elicitation-form_of_response-3": {
    "data": {
        "Informativeness": false,
        "Non-redundancy": false,
        "Fluency": false
    },
    "control": {},
    "text": {
        "Informativeness": "",
        "Non-redundancy": "",
        "Fluency": ""
    }
},
"heds-sample_evaluators_design-sample-statistical_power-method": {
   "data": {
      "": "N/A"
   },
   "control": {}
},
"heds-sample_evaluators_design-sample-system_output_selection-4": {
   "data": {
      "": false
   },
   "control": {},
   "text": {
      "": ""
   }
},
"heds-criteria-criterion-response_elicitation-inter_annotator-agreement-2": {
   "data": {
      "Informativeness": false,
      "Non-redundancy": false,
      "Fluency": false
   },
   "control": {},
   "text": {
      "Informativeness": "",
      "Non-redundancy": "",
      "Fluency": ""
   }
},
"heds-criteria-criterion-response_elicitation-scale_presented_as-4": {
   "data": {
      "Informativeness": false,
      "Fluency": false,
      "Non-redundancy": false
   },
   "control": {},
   "text": {
      "Informativeness": "",
      "Fluency": "",
      "Non-redundancy": ""
   }
},
"heds-paper_and_resources-paper-experiment_identification": {
   "data": {
      "": "Human Evaluation (Section 4.3 and Section 5.2)"
   },
   "control": {}
},
"heds-sample_evaluators_design-evaluators-evaluators-are_authors-3": {
   "data": {
      "": false
   },
   "control": {},
   "text": {
      "": ""
   }
},
"heds-criteria-criterion-response_elicitation-form_of_response-other_text": {
   "data": {
      "Informativeness": "",
      "Non-redundancy": "",
      "Fluency": ""
    },
    "control": {}
  },
  "heds-sample_evaluators_design-evaluators-characteristics": {
    "data": {
      "": "PhD candidates in computer science\n2 males, 1 female"
    },
    "control": {}
  },
  "heds-sample_evaluators_design-evaluators-evaluators-known_to_authors-3": {
    "data": {
      "": false
    },
    "control": {},
    "text": {
      "": ""
    }
  },
  "heds-criteria-criterion-response_elicitation-inter_annotator-agreement-1": {
    "data": {
      "Informativeness": true,
      "Non-redundancy": true,
      "Fluency": true
    },
    "control": {},
    "text": {
      "Informativeness": "1. yes",
      "Non-redundancy": "1. yes",
      "Fluency": "1. yes"
    }
  },
  "heds-criteria-criterion-criteria-self_vs_external_frame-2": {
    "data": {
      "Informativeness": true,
      "Non-redundancy": true,
      "Fluency": false
    },
    "control": {},
    "text": {
      "Informativeness": "2. Quality of output relative to the input",
      "Non-redundancy": "2. Quality of output relative to the input",
      "Fluency": ""
    }
  },
  "heds-ethics-impact_assessments": {
    "data": {
      "": "No"
    },
    "control": {}
  },
  "heds-criteria-criterion-response_elicitation-size_of_scale-1": {
    "data": {
      "Informativeness": true,
      "Non-redundancy": true,
      "Fluency": true
    },
    "control": {},
    "text": {
      "Informativeness": "1. Discrete",
      "Non-redundancy": "1. Discrete",
      "Fluency": "1. Discrete"
    }
  },
  "heds-criteria-criterion-response_elicitation-participant_criterion_definiiton": {
    "data": {
      "Informativeness": "Does the generated summary correctly cover the information in the ground truth summary?\n(标准答案是由多个子句组成的，这里我们想要判断标准答案中的每子句的信息是否被抽取到了。)",
      "Non-redundancy": "Does the generated summary not contain repeated, meaningless or unnecessary information?\n(待测摘要文本也是由多个子句组成的，这里我们想要判断待测文本中的每个子句的信息是否是冗余的。)",
      "Fluency": "Is the generated summary well-formed, semantically complete, and easy to understand?\n(我们想要判断待测文本中的每个子句的语言表达流畅性。)"
    },
    "control": {}
  },
  "heds-sample_evaluators_design-sample-statistical_power-script": {
    "data": {
      "": "N/A"
    },
    "control": {}
  },
  "heds-sample_evaluators_design-experimental_design-experimental_conditions-5": {
    "data": {
      "": false
    },
    "control": {},
    "text": {
      "": ""
    }
  },
  "heds-criteria-criterion-criteria-output_aspect-3": {
    "data": {
      "Informativeness": false,
      "Non-redundancy": true,
      "Fluency": false
    },
    "control": {},
    "text": {
      "Informativeness": "",
      "Non-redundancy": "3. Both form and content of output",
      "Fluency": ""
    }
  },
  "heds-criteria-criterion-response_elicitation-scale_presented_as-1": {
    "data": {
      "Informativeness": false,
      "Fluency": false,
      "Non-redundancy": false
    },
    "control": {},
    "text": {
      "Informativeness": "",
      "Fluency": "",
      "Non-redundancy": ""
    }
  },
  "heds-sample_evaluators_design-experimental_design-evaluators_can_ask_questions-1": {
    "data": {
      "": true
    },
    "control": {},
    "text": {
      "": "1. evaluators are told they can ask any questions during/after receiving initial training/
instructions, and before the start of the evaluation"
    }
  },
  "heds-criteria-criterion-response_elicitation-form_of_response-11": {
    "data": {
      "Informativeness": false,
      "Non-redundancy": false,
      "Fluency": false
    },
    "control": {},
    "text": {
      "Informativeness": "",
      "Non-redundancy": "",
      "Fluency": ""
    }
  },
  "heds-paper_and_resources-names_and_affiliations-person_completing_this_sheet-email": {
    "data": {
      "": "t-ito@tohoku.ac.jp"
    },
    "control": {}
  },
  "heds-criteria-criterion-evaluation_mode-absolute_or_relative-other_text": {
    "data": {
      "Informativeness": "",
      "Non-redundancy": "",
      "Fluency": ""
    },
    "control": {}
  },
  "heds-sample_evaluators_design-experimental_design-experimental_conditions-2": {
    "data": {
      "": false
    },
    "control": {},
    "text": {
      "": ""
    }
  }
}
```