ReproHum #0892-01: The painful route to consistent results: A reproduction study of human evaluation in NLG

Irene Mondella[⋆], **Huiyuan Lai**[◇], **Malvina Nissim**[◇]
[⋆]ILC-CNR / University of Pisa, Italy
[◇]CLCG, University of Groningen, the Netherlands
i.mondella@studenti.unipi.it
{h.lai,m.nissim}@rug.nl

Abstract

In spite of the core role human judgement plays in evaluating the performance of NLP systems, the way human assessments are elicited in NLP experiments, and to some extent the nature of human judgement itself, pose challenges to the reliability and validity of human evaluation. In the context of the larger ReproHum project, aimed at running large scale multi-lab reproductions of human judgement, we replicated the understandability assessment by humans on several generated outputs of simplified text described in the paper "Neural Text Simplification of Clinical Letters with a Domain Specific Phrase Table" by Shardlow and Nawaz, appeared in the Proceedings of ACL 2019. Although we had to implement a series of modifications compared to the original study, which were necessary to run our human evaluation on exactly the same data, we managed to collect assessments and compare results with the original study. We obtained results consistent with those of the reference study, confirming their findings. The paper is complete with as much information as possible to foster and facilitate future reproduction.

Keywords: human evaluation, reproducibility, ReproHum

1. Introduction

Human evaluation of model performance plays a central role in Natural Language Processing (NLP). This is particularly true in the broadly defined area of Natural Language Generation (NLG), which encompasses machine translation, rephrasing, summarisation, etc, i.e., any modelling task whose output consists in some generated text. Indeed, the large variability in acceptable outputs does not allow for an exhaustive set of gold references to be pre-produced, as is instead the case for classification tasks. For the same reason, automatic metrics must be used that are able to capture some degree of similarity between references and different but potentially valid outputs, and cannot exploit an exact correspondence of reference and output.

Developments in NLG evaluation have seen the direct incorporation of human judgements into trainable metrics, such as COMET (Rei et al., 2020), leading to much higher correlations to human assessments. While on the one hand the development of metrics that better align to human judgement appears to be a very promising direction, on the other hand the optimism could be tainted by findings along another avenue of research, dedicated to the *reproducibility* (and therefore reliability) of human judgement.

Recent efforts conducted in the context of the ReproGen shared evaluation campaigns (Belz et al., 2021, 2022) and especially the preliminary findings of ReproHum[1] (Belz et al., 2023), a cooperative project aimed to test the replicability of human evaluations reported in existing NLP papers through large-scale reproductions across multiple research groups, have shed some worrying light on the reliability – and thus validity – of human assessments themselves. Strikingly, Belz et al. (2023, p. 5) report "that only a small fraction of previous human evaluations in NLP can be repeated under the same conditions, hence that their reproducibility cannot be tested by repeating them."

The present paper reports on a reproduction experiment which is also part of the ReproHum project (Belz and Thomson, 2024), as an ongoing effort to further explore the extent to which human judgements elicited in NLP, and in this context more specifically NLG experiments, can be considered reliable and what mostly affects reproduction. As part of ReproHum, our work follows the research template provided by the project coordination team; this paper presents our results accordingly, thus following specific guidelines and reporting templates. The experiment was pre-registered through the Human Evaluation Data Sheet (HEDS[2]) as proposed by Shimorina and Belz (2022). We first introduce the details of the original experiment and the human evaluation included therein, and then describe our own reproduction study, specifically focusing on all the adjustments we had to made in our experiments compared to the original evaluation setup. We compare results critically, running a correlation analysis and comparing inter-annotator agreement across the two studies. We observe that our efforts in faithfully reproducing the original human evalua-

[1] https://reprohum.github.io/

[2] Details at the following link: https://github.com/nlp-heds/repronlp2024

Proceedings of the HumEval @LREC-COLING 2024, pages 261–268
Turin, Italy. May 21st, 2024. ©2024 European Language Resources Association (ELRA)

Complex Term	Simple Term
ability to be ambulant	ability to walk
carcinoma of stomach	cancer of stomach
hypertension	high blood pressure
osteophyte	bony spur

Table 1: Examples of term pairs for phrase table.

tion have, in this case, brought promising results: the findings of the reference paper were confirmed by our reproduction, and the changes we had to make to the original experimental design did not affect the consistency between the two studies.

2. Overview of Original Study

We aim to reproduce the human evaluation experiment of text simplification in "Neural Text Simplification of Clinical Letters with a Domain Specific Phrase Table" by Shardlow and Nawaz (2019). Text simplification is the process of automatically paraphrasing a text to improve its understandability while preserving its original meaning (Al-Thanyyan and Azmi, 2021). This has a wide range of applications, such as helping non-native speakers and bridging the gap between layman and expert.

2.1. Task and Models

This original study aims to use text simplification methods to automatically aid patient understanding of clinical letters containing complex medical terminology (see examples in Table 5). Specifically, based on the SNOMED-CT clinical thesaurus (Donnelly, 2006), the authors created a phrase table that links complex medical terminology to simpler vocabulary (see Table 1), which is used to augment existing neural text simplification systems. To assess the impact of the proposed method on the ease of understanding sentences, human judgment is elicited to evaluate three different systems as well as the original sentences, for a total of four versions of the same sentence:

- **Original Texts (ORIG):** The original texts appear after preprocessing, which ensures that they are equivalent to the transformed texts and that any effects would be from the simplification system, not the preprocessing.
- **NTS:** The original sentences were modified by the Neural Text Simplification (NTS) system (Nisioi et al., 2017), which uses the open-source OpenNMT (Klein et al., 2017) library that provides sequence to sequence learning between a source and target language.
- **NTS + Phrase Table (NTS + PT):** The original sentences were modified by NTS, but when OpenNMT identified a word as being out-of-vocabulary, this system (the one proposed by the authors of the original paper) will use the phrase table to replace it.
- **Phrase Table Baseline (PTB):** To demonstrate the advantages of using the phrase table in tandem with the NTS system, the proposed baseline is to only apply the phrase table to every word that could be replaced in the text.

The simplified sentences, generated by the systems described above, as well as the original version, are assessed by means of human evaluation.

2.2. Human Evaluation

The original study selected 50 source texts from two different datasets: i2b2 (Uzuner et al., 2007), which is a dataset of 899 discharge summaries, and MIMIC-III v1.4 (Johnson et al., 2016), which contains over 58,000 hospital records, with detailed clinical information. In this way, they obtained 100 instances: for each of them, 3 different simplified versions were created using the methods described in subsection 2.1, obtaining 100 4-tuples of parallel sentences. Texts within a 4-tuple are identical except for the modifications made by each system. No two sentences in a 4-tuple are the same.

The human evaluation was conducted on Figure Eight, a crowd-sourcing platform that no longer exist. Each 4-tuple has been assessed by 10 annotators, and each annotator could complete a maximum of 20 annotations, with the aim of obtaining a wide variety of perspectives on the data. No annotator saw the same 4-tuple twice.

To ensure the quality of annotations, workers with a higher than average rating on the Figure Eight platform were selected (level 2 and above), and a set of test annotations was designed to filter out bad-actors. From the analysis of the raw results, we found that there was a total of 8 test annotations, and most of the participants had to answer to all of them.

For each 4-tuples, annotators have been asked to rank the 4 sentences according to their ease of understanding, where the top-ranked sentence (rank 1) is the easiest to understand, while the bottom-ranked sentence (rank 4) is the hardest. Furthermore, it was specified that, in the case of 2 sentences of equal complexity, the annotator should order them according to the order of presentation. In total, 1000 annotations (100 instances with 10 annotations each) were collected. However, 20 of them were identified as not using all 4 ranks, i.e. 2 or more sentences were at the same ranking level. In these cases, the specific annotation was removed in the final analysis, resulting in 980 rankings.

Setting	Original Study	Replicated Study
Platform	Figure Eight	Prolific
Participants	98	40
Conditions	$\geq$level 2	acceptance rate$\geq$99% & completed tasks $\geq$ 200 region filter: UK, USA, Australia, Canada
Filtering	a set of test annotations	3 additional test annotations
Reward	Unknown	£12 per hour

Table 2: Human evaluation settings in original and replicated study.

Finally, the authors design a metric to calculate the average rank r_s of a system s, which is described in Equation 1.

$$r_s = \frac{\sum_{i=1}^{4} i \times f(s,i)}{\sum_{i=1}^{4} f(s,i)} \quad (1)$$

where i is a rank from 1 to 4 and $f(s,i)$ is a function that maps the system and rank to the number of times that system is placed at that rank.

3. Reproduction Study

In our reproduction study, we strictly followed the settings of the human evaluation performed by the authors of the original work, although some adjustments had to be made for various reasons.

First, we couldn't use the crowd-sourcing platform used in the original study, because it doesn't exist anymore, so we used instead Prolific[3]. One of the main differences between these two platforms is that in Prolific it is necessary to set in advance the number of items to be evaluated by each participant. Analysing the raw results of the original paper, we assume that this constraint was not present in Figure Eight, since 76 participants evaluated 20 4-tuples (the maximum number set by the authors of the original study, included the test annotations), and 22 participants rated fewer items. In total, in the original evaluation, 98 annotators were recruited. In our case, however, it was necessary to create surveys of a fixed length. To conform our reproduction to the experimental design adopted in the ReproHum project, we created surveys containing 25 instances. Also, to ensure quality of annotations, we added 3 additional test annotations to each survey to filter out bad actors. Since the total number of instances is 100, this resulted in 4 different surveys, each of them presented to 10 different participants, for a total of 40 annotators. We made sure that no annotator participated in more than one survey.

[3] https://www.prolific.com/

Another difference in our replication, made necessary by the use of a different crowd-sourcing platform, regards the selection criteria for participants. Since we do not know how the participants' rating was calculated in Figure Eight, we opted to set, on Prolific, a minimum acceptance rate of 99% and a minimum completed tasks of 200. In addition, we saw from the original raw results that all the evaluators were in the United Kingdom, United States, or Australia. Whether this is by design or not we cannot tell for sure; however, because of this strong evidence, we set a region filter on English-speaking countries UK, US, Australia and Canada.

Another point on which we acted independently of the original experiment is the compensation due to the annotators, not specified in the reference paper. In our reproduction, we followed the current UK minimum wage of £12 per hour, following the general recommendation of the ReproHum project. Estimating a minimum completion time of 30 minutes per survey, we payed £6 per participant.

The differences in settings between the human evaluation performed by us and the original one are summarized in Table 2.

A screenshot of the annotation interface we created is shown in Figure 1, with instructions reported also as a screenshot in Figure 2 (the latter in the Appendix). Instead of creating a different question for each of the 4 sentences, as in the original annotation interface, we opted for a drag-and-drop system, that we find more intuitive. The instruction page, on the other hand, is faithfully copied from the original (excluding the parts explaining how to answer questions, for which we have adapted the instructions to our annotation interface).

4. Results

One of the main difficulties one faces in faithfully reproducing an experiment carried out by others lies in gathering all the necessary information. If they are not directly stated in the reference paper, it is necessary to seek clarification from the persons involved. However, during this exchange of infor-

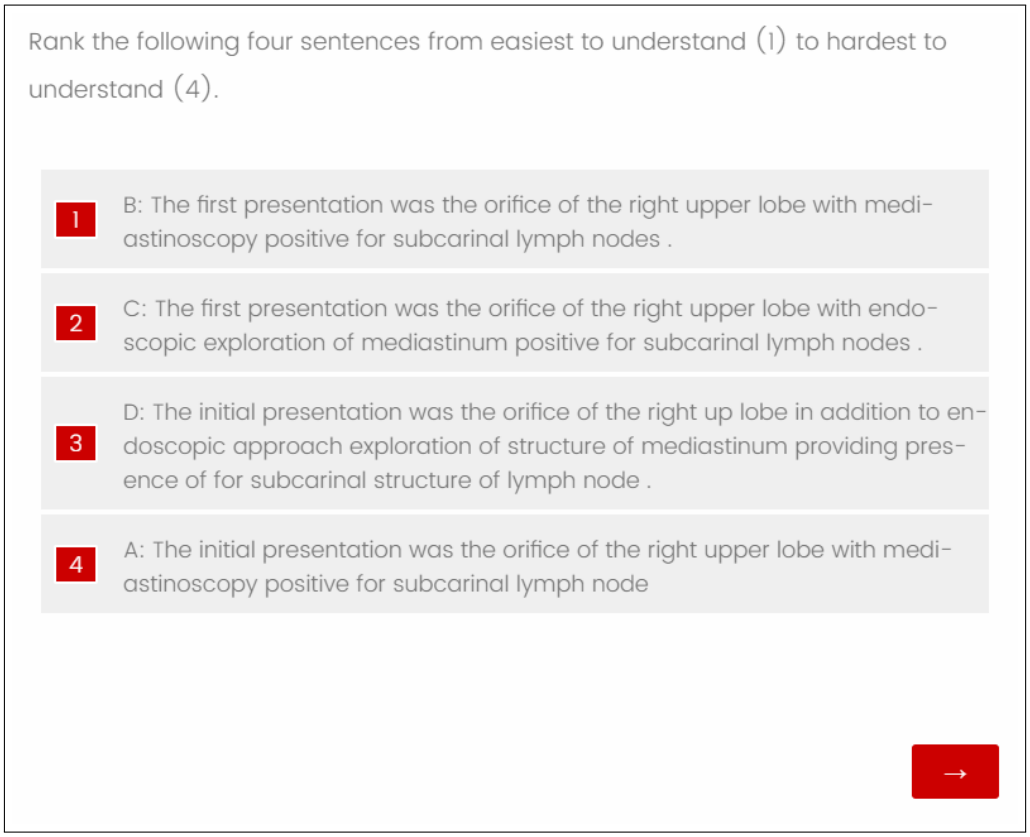

Figure 1: A screenshot of the annotation interface used in our replication study

System	Rank:1		Rank:2		Rank:3		Rank:4		AVG		CV*
	O	R	O	R	O	R	O	R	O	R	
NTS + PT	430	517	255	214	230	197	65	72	**1.93**	**1.82**	5.63
NTS	259	228	294	288	264	276	163	208	2.34	2.46	5.15
ORIG	120	123	222	233	381	408	257	236	2.79	2.76	1.19
PTB	171	132	209	265	105	119	495	484	2.94	2.96	0.51

Table 3: Comparison of original and reproduced results. *Rank:x* indicates the number of times each system was ranked at rank *x* and the last two columns show the average rank calculated according to the formula 1. O = Original and R = Reproduced. CV* is the Coefficient of Variation.

mation and material, doubts or misunderstandings may arise, as happened in this case: the results we are now going to present were, initially, completely different, due to a wrong assignment of the outputs to the 4 systems analysed. We find it interesting to mention this incident, as it is the consequence of one of the inherent difficulties of a reproducibility study such as this one.

Side-by-side Results Table 3 reports comparative results for the original (O) and reproduced (R) studies. It should be noted that the total number of annotations taken into account in the final results varies between the original experiment and our replication. This is due to the fact that, as mentioned in subsection 2.2, the authors of the original study had to remove 20 annotations, resulting in 980 final data points. In the reproduced results shown in Table 3, however, no annotations was removed (resulting in 1000 final data points), because all of them meet the response criteria.

What emerges from our study confirms the original results: the system proposed by the authors of the reference paper (NTS + PT) is the best performing one in their case, with an average rank of 1.93, and it is also the best one in our reproduction (1.82). Moreover, the general order of all systems turns out to be the same, with the Phrase Table Baseline as the worst performing one, generating outputs that are, in average, less understandable than the original sentences.

Reproducibility Analysis Following the protocol for the ReproHum project, in Table 3 we reported

	Krippendorff's α	Pearson's r	Spearman's ρ
Agreement between Two Studies	0.30	-	-
IAA of Original Study	0.22	-	-
IAA of Replicated Study	0.40	-	-
Corr. between Two Studies (System Scores)	-	0.98	1.00
Corr. between Two Studies (Average Annotations)	-	0.76	0.75

Table 4: Agreement between the two studies, calculated considering all 20 annotations for each sentence; IAA for the original and the replication study; correlation coefficients between the two experiments' results; correlation coefficients between the two experiments' sets of average annotations.

System	Sentence	O	R
ORIG	A diagnostic paracentesis was said to show a sterile transudate.	2.9	3.6
NTS	A diagnostic paracentesis was said to show a good transudate.	2.2	2.5
NTS + PT	A diagnostic puncture and drainage was said to show a good transudate.	1.3	1.1
PTB	A diagnostic has intent puncture and drainage was said to show a sterile transudate.	3.6	2.8
ORIG	The tumor now involves the trachea as well as the right main bronchus.	2.8	2.1
NTS	The tumor now involves the opening as well as the right main bronchus.	2.0	1.9
NTS + PT	The tumor now involves the opening as well as the right main bronchial structure.	1.5	2.3
PTB	The tumor now involves the tracheal structure as good as the right main bronchial structure.	3.7	3.7

Table 5: Examples of outputs produced by different systems and corresponding results from the original (O) and reproduced (R) rankings.

the Coefficient of Variation debiased for small sample size (CV*), as defined in Belz (2022).

We then calculated the agreement between ours and the original results, by considering all 20 annotators (10 from the original experiment and 10 from our reproduction study) for each sentence. We used the Krippendorff's α agreement measure as proposed in Castro (2017), and achieved an agreement of 0.30, as shown in Table 4. In the same table, we also reported the Inter-Annotator Agreement both within the evaluations collected by us and those collected by the authors of the reference study, for which we achieved higher scores.

We also calculated the correlation between the two sets of system final scores: ours and the original one, as reported in the column "AVG" of Table 3. Table 4 shows that a very high positive correlation was found, consistent with our similar results. To get more information on the quality of our reproduction, however, we also analysed the correlation between the two sets of single evaluations given by our annotators and the evaluations gathered in the original study. Specifically, we assigned each of the 400 annotated sentences (4 sentences for 100 instances) the average score received by the 10 annotators, and ran the correlation between the two studies. The results show that they have high correlation scores on both levels, confirming our results consistent with the original study. Lastly, we reported an error count on these two lists of average rank, rounding the average rank to the nearest whole number, and found that 250 of the 400 values from the two studies agree, while 150 values differ.

Case Study Table 5 shows two examples of annotations: for each example, we reported the four evaluated outputs and the average score obtained by the ten annotators, both in the original experiment and in our reproduction. It can be seen that the NTS + PT system makes targeted changes to the original sentence, managing to modify too technical terms. In the first example, these changes result in increased understandability from the original sentence; however, for the second example, our annotators found the original sentence to be slightly more understandable. The baseline, on the other hand, makes a substantial number of changes, but these do not always help to increase the understandability of the sentence.

5. Conclusion

The main objective of this study was to remain as faithful as possible to the experimental choices made by the authors of the original paper when replicating the human evaluation they ran on system outputs. Any independent decisions we made were motivated by contingencies beyond our con-

trol (such as the use of a different crow-sourcing platform) or by a lack of information (e.g., concerning the compensation due to the annotators). Although the reproducing process is intrinsically difficult, the results we obtained align with the general findings of the original paper.

Acknowledgments

We are very grateful to the anonymous reviewers for their useful comments, which strengthened this paper. We also thank the annotators for helping us evaluate the data. Finally, we thank the ReproHum project coordinators for their consistent support as we performed the replication experiment.

6. Bibliographical References

Suha S. Al-Thanyyan and Aqil M. Azmi. 2021. Automated text simplification: A survey. *ACM Comput. Surv.*, 54(2).

Anya Belz. 2022. A metrological perspective on reproducibility in nlp. *Computational Linguistics*, 48(4):1125–1135.

Anya Belz, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. 2021. The ReproGen shared task on reproducibility of human evaluations in NLG: Overview and results. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 249–258, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Anya Belz, Anastasia Shimorina, Maja Popović, and Ehud Reiter. 2022. The 2022 ReproGen shared task on reproducibility of evaluations in NLG: Overview and results. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 43–51, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Anya Belz and Craig Thomson. 2024. The 2024 repronlp shared task on reproducibility of evaluations in nlp: Overview and results. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*.

Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Anouck Braggaar, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondřej Dušek, Steffen Eger, Qixiang Fang, Mingqi Gao, Albert Gatt, Dimitra Gkatzia, Javier González-Corbelle, Dirk Hovy, Manuela Hürlimann, Takumi Ito, John D. Kelleher, Filip Klubicka, Emiel Krahmer, Huiyuan Lai, Chris van der Lee, Yiru Li, Saad Mahamood, Margot Mieskes, Emiel van Miltenburg, Pablo Mosteiro, Malvina Nissim, Natalie Parde, Ondřej Plátek, Verena Rieser, Jie Ruan, Joel Tetreault, Antonio Toral, Xiaojun Wan, Leo Wanner, Lewis Watson, and Diyi Yang. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Santiago Castro. 2017. Fast Krippendorff: Fast computation of Krippendorff's alpha agreement measure. https://github.com/pln-fing-udelar/fast-krippendorff.

Kevin Donnelly. 2006. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279–90.

Alistair Johnson, Tom Pollard, Lu Shen, Li-wei Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Celi, and Roger Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Matthew Shardlow and Raheel Nawaz. 2019. Neural text simplification of clinical letters with a domain specific phrase table. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 380–389, Florence, Italy. Association for Computational Linguistics.

Anastasia Shimorina and Anya Belz. 2022. The human evaluation datasheet: A template for recording details of human evaluation experiments in NLP. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 54–75, Dublin, Ireland. Association for Computational Linguistics.

Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the State-of-the-Art in Automatic De-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563.

A Appendix

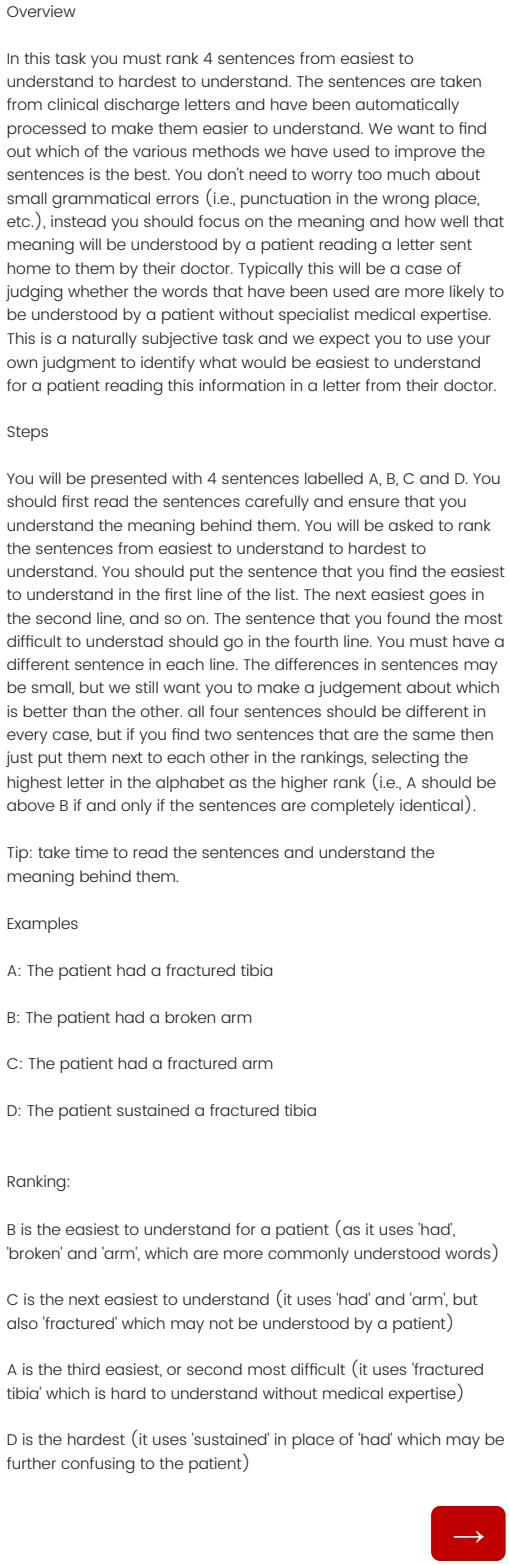

Figure 2: A screenshot of the instruction interface in our replication study.