Reproducing a comparison of hedged and non-hedged NLG texts

Saad Mahamood
trivago N.V.
Düsseldorf, Germany
saad.mahamood@trivago.com

Abstract

This paper describes an attempt to reproduce an earlier experiment, previously conducted by the author, that compares hedged and non-hedged NLG texts as part of the ReproGen shared challenge. This reproduction effort was only able to partially replicate results from the original study. The analysis from this reproduction effort suggests that whilst it is possible to replicate the procedural aspects of a previous study, replicating the results can prove more significantly challenging as differences in participant type can have a potential impact.

1 Introduction

There has been within recent years a great interest in understanding and quantifying the reproducibility of experiments across several areas of scientific research. This also includes experiments in the field of Natural Language Understanding (NLU), where researchers have questioned the degree to which experiments and results can reliably be reproduced. Recent working exploring the reproducibility of past NLU work has found significant issues such as only a minority of systems reproducing previously reported scores and systems not working due to non-functional code or resource limits (Belz et al., 2021). Additionally, there has been growing awareness of systematic issues with regards to how human evaluations are being conducted. In particular, the lack of standardisation and significant under reporting of key human evaluation details (Howcroft et al., 2020). These twin concerns has led to the creation of the ReproGen shared task (Belz et al., 2020), which attempts to check the reproducibility of human evaluations within the field of Natural Language Generation (NLG).

As part of the ReproGen shared task[1], a reproduction experiment was attempted for a previous work that the author had previously conducted in 2007. In this previous work a human evaluation was conducted between NLG texts containing hedge phrases and those that do not (Mahamood et al., 2007). This past experiment was conducted to better understand the impact of hedge phrases can have when introduced into a data-to-text NLG system. This was done in order to understand how such systems should communicate potentially emotionally sensitive information to a given reader.

In this paper we will describe the experimental setup used and the differences that were made in the reproduction experiment (Section 2), the results obtained and how they compare to the ones originally obtained (Section 3), and finally we will discuss the significance of the results obtained in this reproduction effort (Section 4).

2 Experimental Setup & Differences

2.1 Procedure

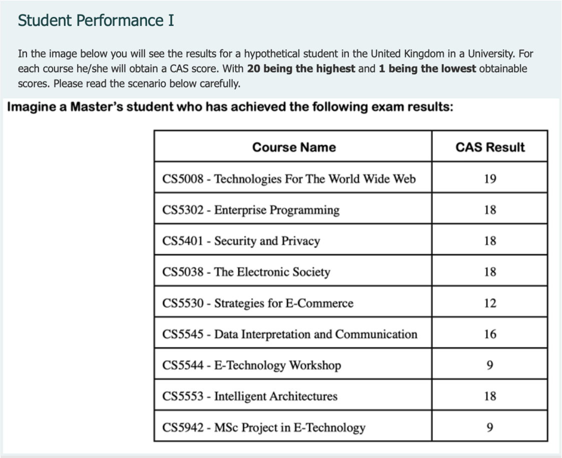
Like the previous experiment, this reproduction experiment sought to obtain individual preferences of participants when presented with hedged and non-hedged texts when communicating exams results for hypothetical exams results. This was done across two differing scenarios. The first in a positive context where a hypothetical strong student has obtained a high set of results as shown in Figure 1.

The second in a negative context where a weak student has obtained a low set of exam results. For each of the scenarios the participants are shown the raw exam scores attained and two texts summarising these results: one with hedges and one without as shown in Figures 2 and 3. In total participants were expected to evaluate four different texts. Two for each scenario with one participant judgement expected per scenario.

Whilst the original experiment was conducted with a paper based questionnaire sheet, the repro-

[1] ReproGen - https://reprogen.github.io/

Proceedings of the 14th International Conference on Natural Language Generation (INLG), pages 282–285,
Aberdeen, Scotland, UK, 20-24 September 2021. ©2021 Association for Computational Linguistics



Student Performance I

In the image below you will see the results for a hypothetical student in the United Kingdom in a University. For each course he/she will obtain a CAS score. With **20 being the highest** and **1 being the lowest** obtainable scores. Please read the scenario below carefully.

Imagine a Master's student who has achieved the following exam results:

Course Name	CAS Result
CS5008 - Technologies For The World Wide Web	19
CS5302 - Enterprise Programming	18
CS5401 - Security and Privacy	18
CS5038 - The Electronic Society	18
CS5530 - Strategies for E-Commerce	12
CS5545 - Data Interpretation and Communication	16
CS5544 - E-Technology Workshop	9
CS5553 - Intelligent Architectures	18
CS5942 - MSc Project in E-Technology	9

Figure 1: High exam results table for the first scenario.

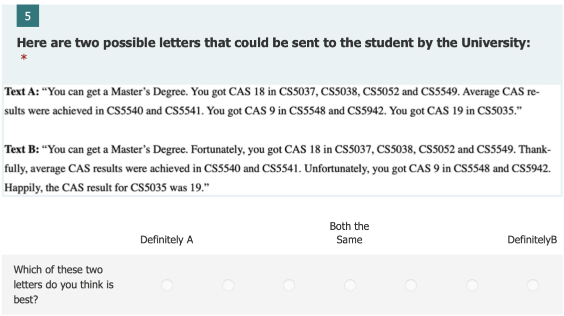
duction used an online based form instead. However, both the questions asked and the format used were mostly identical between the two experiments. The two minor differences being the introduction of additional gender options and the use of age ranges instead of asking participants directly their age.

Participants were asked initially to give their background information. This consisted of their gender (*male*, *female*, *non-binary*, *other*), select an appropriate age-range band, and finally degree of English language proficiency (*Native*, *Non-native, but fluent*, *Not fluent*). Then for both scenarios they were asked to read the results for the student as presented in a table (Figure 1). After this, the participants were asked to state whether they felt the results were good or not for the student (*Yes*, *No*, *Maybe*) and a preference between the two texts A and B. This was done using a Likert scale which ran from -3 for Text A to +3 for Text B. If both texts were considered by a participant to be the same then a score of 0 was given. The participants were asked to provide free text comments on why they made their particular choice of text.

2.2 Participants

The original experiment recruited 37 Masters students (9 females and 28 males). Out of these students only responses from 32 students were used due to incomplete responses from 5 students. From the remaining students 14 participants identified as native English speakers, 11 as non-native but fluent, and 7 as non-fluent English speakers.

Table 1 gives a direct comparison of the participants recruited for the original and reproduction experiments. In contrast the cohort recruited for the reproduction experiment consisted of colleagues from the author's institution. A total of 11 participants were recruited (4 females and 7 males). Five participants identified themselves as fluent native English speakers and six as non-native but fluent English speakers. No non-fluent English speakers were recruited due to the fact that such participants were not available. Additionally, another key difference between the original experiment and the reproduction is the age of the participants. In the original study 44% (*n=15*) of the participants were under 25 years old, whereas in the reproduction experiment only one participant recruited was in this particular age bracket.



5

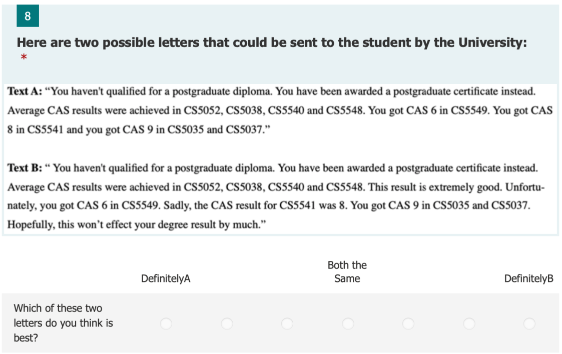
Here are two possible letters that could be sent to the student by the University: *

Text A: "You can get a Master's Degree. You got CAS 18 in CS5037, CS5038, CS5052 and CS5549. Average CAS results were achieved in CS5540 and CS5541. You got CAS 9 in CS5548 and CS5942. You got CAS 19 in CS5035."

Text B: "You can get a Master's Degree. Fortunately, you got CAS 18 in CS5037, CS5038, CS5052 and CS5549. Thankfully, average CAS results were achieved in CS5540 and CS5541. Unfortunately, you got CAS 9 in CS5548 and CS5942. Happily, the CAS result for CS5035 was 19."

	Definitely A			Both the Same			DefinitelyB
Which of these two letters do you think is best?	○	○	○	○	○	○	○

Figure 2: Positive student scenario Text A (without hedges) and B (with hedges).



8

Here are two possible letters that could be sent to the student by the University: *

Text A: "You haven't qualified for a postgraduate diploma. You have been awarded a postgraduate certificate instead. Average CAS results were achieved in CS5052, CS5038, CS5540 and CS5548. You got CAS 6 in CS5549. You got CAS 8 in CS5541 and you got CAS 9 in CS5035 and CS5037."

Text B: " You haven't qualified for a postgraduate diploma. You have been awarded a postgraduate certificate instead. Average CAS results were achieved in CS5052, CS5038, CS5540 and CS5548. This result is extremely good. Unfortunately, you got CAS 6 in CS5549. Sadly, the CAS result for CS5541 was 8. You got CAS 9 in CS5035 and CS5037. Hopefully, this won't effect your degree result by much."

	DefinitelyA			Both the Same			DefinitelyB
Which of these two letters do you think is best?	○	○	○	○	○	○	○

Figure 3: Weak student scenario Text A (without hedges) and B (with hedges).

3 Reproduction Results

The results from the reproduction experiment along with the original experiment results for the native and fluent English speaker groups are shown in Table 2. Since there were no non-fluent English speakers recruited results for only native and non-fluent speaker groups are shown. The biggest difference between from the original and reproduction experiments is the results for fluent speakers of English. In the original study this group had shown a

	Native Speakers	**Non-Native, but Fluent**	**Non-Fluent**	**Total**
Original Study	14 (Male: 11, Female: 3)	11(Male: 9, Female: 2)	7 (Male: 3, Female: 4)	32
Repro. Study	6 (Male: 5, Female: 1)	5 (Male: 2, Female: 3)	0	11

Table 1: Comparison of participant numbers between the original and reproduction studies.

	Native	**Fluent**
Original: S1	-1.42 (σ 2.39)	0.09 (σ 2.59)
Original: S2	-2.07 (σ 1.25)	0.45 (σ 2.53)
Repro: S1	-2.2 (σ 1.09)	-2.0 (σ 1.09)
Repro: S2	-1.40 (σ 2.07)	-1.33 (σ 1.86)

Table 2: Results from the original and reproduction studies for native and fluent speakers. S1 or S2 refers to a particular scenario.

weak preference for hedge texts on average, However, in the reproduction this group like the native speakers show an overall strong preference for non-hedged texts in both scenarios. This difference could potentially be explained by difference in the type of participants (Master students vs. working professionals) recruited between the two studies.

For native speakers, the results of the reproduction confirm the initial findings that native speakers prefer the non-hedged over the hedged texts. Interestingly, like the original study native speakers tend to prefer the non-hedged texts to a higher degree than compared to fluent speakers. Although this effect is less pronounced than compared to the original study.

A two-sample T-test was conducted to compare the mean rating score of the native and fluent speaker groups for both scenarios[2]. For the first scenario the result was *t(9)=-0.301*, *p=0.769* and for the second scenario it was *t(9)=-0.056*, *p=0.956*. The statistically non-significant p-values for both scenarios indicate that the mean rating scores given by both groups for each scenario are not statistically different from each other.

Analysis of free-text comments from fluent speakers across both scenarios showed that participants found the hedges “didn’t add value” and that the non-hedged texts were more “formal” and “professional”. These comments align with the general comments from native speakers from the original study. It is possible that the use of fluent speakers with professional experience of using English results in cultural expectations that are closer to that of native speakers than compared to fluent speaking students of the original study. Therefore the need for hedges to act as “emotional navigators” are significantly diminished for non-native fluent speakers.

4 Conclusion

In this paper we have conducted a reproduction of a previous NLG study. Unfortunately, we have only been able to only partially replicate the results from the original study. Whilst, we were able to confirm the findings for native speakers we were not able to do so for fluent speakers. This suggest two things. Firstly, that reproduction is a necessary step to better understand the validity of results obtained in initial experiments. And until those results have been validated by a reproduction effort such results should be taken with a degree of scepticism. The second key point is that results obtained in earlier studies cannot be generalised beyond a particular target group of human participants until a reproduction effort confirms the same effect with a different audience. In the case of this study, the original experiment was conducted with Master students. It is possible the effects found maybe limited to that audience in particular. Therefore, it is critical that key demographic information is recorded in human evaluiations to enable future reproduction efforts to have the correct participant mix for their experiments.

The two key limitations of this reproduction effort is the differences in participant types and the lack of non-fluent English speakers recruited for the study. Therefore, due to the second limitation in particular, it was not possible to confirm or reject a key claim from the previous study that non-fluent speakers prefer texts that contain hedge phrases. This remains an area open for a possible future follow-up reproduction effort.

References

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2020. ReproGen: Proposal for a shared task on reproducibility of human evaluations in NLG. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages

[2] Reproduction experiment data and analysis code - https://github.com/Saad-Mahamood/reprohum2021

232–236, Dublin, Ireland. Association for Computational Linguistics.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 381–393, Online. Association for Computational Linguistics.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Saad Mahamood, Ehud Reiter, and Chris Mellish. 2007. A comparison of hedged and non-hedged NLG texts. In *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)*, pages 155–158, Saarbrücken, Germany. DFKI GmbH.