Intrinsic Task-based Evaluation for Referring Expression Generation

Guanyi Chen[♣], **Fahime Same**[♡], **Kees van Deemter**[♠]
[♣]Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning,
National Language Resources Monitoring and Research Center for Network Media,
School of Computer Science, Central China Normal University
[♡]Trivago N.V., [♠]Department of Information and Computing Sciences, Utrecht University
g.chen@ccnu.edu.cn, fahime.same@trivago.com, c.j.vandeemter@uu.nl

Abstract

Recently, a human evaluation study of Referring Expression Generation (REG) models had an unexpected conclusion: on WEBNLG, Referring Expressions (REs) generated by the state-of-the-art neural models were not only indistinguishable from the REs in WEBNLG but also from the REs generated by a simple rule-based system. Here, we argue that this limitation could stem from the use of a purely ratings-based human evaluation (which is a common practice in Natural Language Generation). To investigate these issues, we propose an intrinsic task-based evaluation for REG models, in which, in addition to rating the quality of REs, participants were asked to accomplish two meta-level tasks. One of these tasks concerns the referential success of each RE; the other task asks participants to suggest a better alternative for each RE. The outcomes suggest that, in comparison to previous evaluations, the new evaluation protocol assesses the performance of each REG model more comprehensively and makes the participants' ratings more reliable and discriminable.

1 Introduction

Referring Expression Generation (REG) is a key aspect of Natural Language Generation (NLG) and a vital step in the classic Natural Language Generation pipeline (NLG, Reiter and Dale, 2000; Gatt and Krahmer, 2018). REG produces referring expressions (REs) that refer to referents at different points in a discourse (Belz and Varges, 2007). It has great practical value for commercial NLG (Reiter, 2017) and is actively studied by theoretical linguists and psycholinguists (van Deemter, 2016).

Conventional REG systems have two steps. The first step determines the rough form of the RE. For instance, given a certain context, the system needs to decide whether a given reference to Homer Simpson should be a proper name, a pronoun, or a description. The second step involves choosing a concrete Noun Phrase. For example, if step 1 says a proper name must be chosen, then the second step involves choosing between "*Homer*" or "*Homer Simpson*"; if a description must be chosen, then many options exist, including "*the protagonist of the American animated sitcom The Simpsons*".

Castro Ferreira et al. (2018a) redefined the task of REG on the WebNLG corpus (Gardent et al., 2017a; Castro Ferreira et al., 2018b), which has been widely used in NLG research, and tackled it in an End2End manner, i.e. addressing the two above-mentioned steps simultaneously using Deep Neural Networks. Later, many follow-ups were carried out to strengthen the "NeuralREG" model (e.g. Cao and Cheung (2019); Cunha et al. (2020)).

Nonetheless, in a large-scale human evaluation of NeuralREG models, Same et al. (2022) surprisingly found that, for human readers, the REs generated by NeuralREG models are not only indistinguishable from the REs in a corpus but also from the REs generated by a simple rule-based model. For this reason, they questioned the usefulness of neural models in REG, as well as the suitability of the WebNLG corpus as a testing ground for the strengths of the models.

An alternative take on these counter-intuitive results is that the method of asking participants to rate models' outputs may not have been sufficiently sensitive. As previous studies have pointed out (Thomson et al., 2023; Goyal et al., 2023), pragmatic appropriateness is difficult to assess, especially for REs, since the use of REs is highly context-dependent and diverse. Note that previous human evaluations for REG models (Castro Ferreira et al., 2018a; Cunha et al., 2020; Same et al., 2022) always asked participants to rate an entire discourse, which may have failed to make participants aware of the use of REs (even though their instructions asked them to focus on REs).

In this paper, we use Intrinsic Task-based Human Evaluation. In addition to rating each dis-

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7220–7231

Triples: AWH_Engineering_College | country | India Kerala | leaderName | Kochi AWH_Engineering_College | academicStaffSize | 250 AWH_Engineering_College | state | Kerala AWH_Engineering_College | city | "Kuttikkattoor" India | river | Ganges
Text: AWH Engineering College is in Kuttikkattoor, India in the state of Kerala. The school has 250 employees and Kerala is ruled by Kochi. The Ganges River is also found in India. **Delexicialised Text**: AWH_Engineering_College is in "Kuttikkattoor" , India in the state of Kerala . **AWH_Engineering_College** has 250 employees and Kerala is ruled by Kochi . Ganges is also found in India .

Table 1: An example data from the WEBNLG corpus. In the delexicalised text, every entity is underlined, the target entity is **boldfaced**, the pre-context is coloured in blue, and the post-context is coloured in red. The upper part of the table shows the 6 predicate-argument relations; therefore, the size of the triple is 6.

course, participants were asked to accomplish two meta-level tasks. One task is judging the *referential success of each RE*, i.e. determining whether the RE indeed refers to its intended target referent. The other involves *optimising the REs* that, in their opinion, are not optimal for enhancing the clarity and coherence of the discourse.

We believe that this new approach to evaluation has much to add to existing approaches because (1) answering the questions about REs and optimising REs by hand helps participants zoom in on the use of REs in each discourse during the human evaluation, making their score more reliable and discriminable; (2) the participants' responses allow us to evaluate REG models in a more comprehensive way and from multiple angles; and (3) on the basis of rewritings from participants, we can gain further insights about human language use.

In the current study, based on the above protocol, we carried out a human evaluation experiment on the models' outputs from Same et al. (2022). Building on participants' responses, we first analysed the performance of each REG model comprehensively; we then compared our results to those of Same et al. (2022) to see how accomplishing the above-mentioned tasks has impacted participants' rating behaviour. Finally, we looked into the contents of these responses and summarised our observations.

2 Background

First, we explain the REG task and the corpus it is based on. Then, we outline the experiments of Same et al. (2022).

2.1 The REG task

Given a text whose REs are not yet realised, the REG task is to generate the REs that refer to their intended referents. This definition was further elaborated by Castro Ferreira et al. (2018a) based on the WEBNLG corpus.

WEBNLG was constructed by asking crowdworkers to write descriptions for a set of Resource Description Framework (RDF) triples (Table 1); each entity is represented by its delexicalised proper name (e.g. "AWH_Engineering_College" for the entity AWH Engineering College); we call this representation the *identifier* of the entity as it is unique in the corpus. The number of triples varied from 1 to 7. To fit the REG task, Castro Ferreira et al. (2018a) used triples to delexicalise each description (see the delexicalised text in Table 1) and trained a model to generate the surface form of the entity (i.e. *The school*) given its identifier (i.e. "AWH_Engineering_College"), its pre-context ("*AWH_Engineering_College is in "Kuttikkattoor" , India in the state of Kerala .*"), and its post-context ("*has 250 employees and Kerala is ruled by Kochi . The Ganges River is also found in India.*"). To test the generalizability of REG models, Castro Ferreira et al. (2018a) divided the test set of WEBNLG into *seen* set (where all data are from the same domains as the training data) and *unseen* set (where all data are from different domains than the training data).

2.2 Same et al. (2022) Original Findings

Based on the WEBNLG corpus and the REG task defined by Castro Ferreira et al. (2018a), Same et al. (2022) evaluated several state-of-the-art (SOTA) neural REG models, including `ATT+Copy` (Castro Ferreira et al., 2018a), `ATT+Meta` (Cunha et al., 2020) and `ProfileREG` (Cao and Cheung, 2019). Additionally, they tested two rule-based models, one with a small set of rules (`RREG-S`) and one with a larger set of rules (`RREG-L`); as well as two machine learning-based (ML) models, one with a

small set of features (ML-S) and one with a large set of linguistically motivated features (ML-L).[1]

Counter-intuitively, Same et al. found out that in the case of WebNLG, participants found all models to be almost indistinguishable from each other and from the REs in the original WEBNLG documents in terms of clarity, grammaticality, and coherence. They hypothesised that this might be because the data in WEBNLG does not accurately reflect the everyday use of REs. Consequently, they built a new REG dataset based on the Wall Street Journal (WSJ) portion of the OntoNotes corpus (Weischedel, Ralph et al., 2013). More surprisingly, assessments on WSJ indicated that both the simple rule-based (RREG-S) model and the linguistically-informed ML-based (ML-L) model significantly outperformed the two more advanced neural REG models in terms of human evaluation.

These unexpected results, coupled with the lack of distinguishable statistically significant differences in the case of the WEBNLG models, motivated us to perform a more in-depth evaluation of the model outcomes. We begin this evaluation with the simpler corpus, namely, WEBNLG, and plan to expand it in the future to the more complex dataset.

3 Research Questions

As motivated in the introduction, our enhanced experiment rests on three key tasks. The first of these was the same ratings-based task that had been used in earlier studies (including Same et al. 2022), namely, to rate the clarity, grammaticality, and coherence of the item in which a given coreference chain occurs (see Figure 1). The second task was to judge the referential success of the REs. The third was for participants to suggest a rewriting for each RE wherever they felt it was necessary.[2] Our principal aim was to find out how the outcomes of the new experiment differ from earlier experiments; in particular, we were curious *whether the indistinguishability results of Same et al. (2022) may have been caused by the limitations of a human evaluation method that is based solely on ratings.*

Since our experiment produced three different ways of assessing the quality of an RE (namely its rating, its referential success, and the frequency of rewriting), the question comes up about *how these three assessments are related to each other*: do they tend to point in the same direction or not? Note that even if these assessments proved to be very closely aligned, it would not follow that two of the three are superfluous because the fact that this plurality of tasks forced participants to immerse themselves deeply in the texts may have improved the quality of all their responses, including those for the familiar ratings-based task.

Finally, by studying the corpus, it occurred to us that referents differed sharply from each other along a potentially important dimension, namely whether a participant was previously familiar with the referent or not. It has been pointed out that, in some settings, both speakers and hearers use/interpret REs differently if they are familiar with the referent (Kutlak et al., 2011; Staliūnaitė et al., 2018). We, therefore, asked participants, for each referent, whether or not they were familiar with this referent before they read the text, and we investigated *how participants' familiarity with a referent affected their responses on the key tasks*.

Based on the assumptions mentioned above and the questions raised by these assumptions, we put forward the following hypotheses:

Hypothesis 1 ($\mathcal{H}_1$) The meta-level tasks (i.e. the referential success task and rewritings) help participants make more informed ratings. Therefore, unlike the findings by Same et al. (2022), we expect to observe significant distinguishable differences in the ratings of the models.

Hypothesis 2 ($\mathcal{H}_2$) Since the tasks help participants identify inappropriate REs in each text, we expect that the scores in this experiment would be lower than those in Same et al. (2022).

Hypothesis 3 ($\mathcal{H}_3$) Regarding the referential success of REs, we expect that the more frequently REs are marked as successful, the higher the discourse would be rated.

Hypothesis 4 ($\mathcal{H}_4$) We expect that the more often the REs are re-written, the lower the scores the discourse would receive.

Hypothesis 5 ($\mathcal{H}_5$) We expect that participants would spot more inappropriate REs in dis-

[1] Since the underlying mechanisms of these models are not the focus of the present paper, we skip the introduction of their details. Please check Same et al. (2022) for details.

[2] Note that similar ideas were used in the machine translation evaluation (Bentivogli et al., 2018) and the extrinsic evaluation of practical NLG systems (in contrast to our intrinsic evaluation), such as in Sripada et al. (2005), where post-edits were used to understand how experts think about the outputs of a weather forecast generation system.

course if they are familiar with its major entity and, as a consequence, they would re-write more and rate the discourse lower.

4 Intrinsic Task-based Human Evaluation

Here, we describe the setup of the new experiment. We consider this to be an "intrinsic" NLG evaluation, as opposed to the non-real-world ("extrinsic") tasks more commonly associated with task-based NLG evaluation (Reiter, 2011).

4.1 Materials

We used the same set of items from the human evaluation experiment of Same et al. (2022), in which there are seen data and unseen data sampled from the test set of WEBNLG.[3] More specifically, there are 48 seen items (4 items from each triple size group of 2-7) and 48 unseen items (6 items from each triple size group of 2-5). In this study, we considered 5 REG models. Based on models' performance on both WEBNLG and WSJ reported in Same et al. (2022), we selected the two best-performing neural models (`ATT+Copy` and `ATT+Meta`, henceforth, `NREG-1` and `NREG-2`, respectively, since the specific details of each model are not our focus), the best rule-based model (`RREG-S`, henceforth, `RULE`), and the best ML-based model (`ML-L`, henceforth, `ML`)[4]. In other words, for each test item, in addition to the reference text, we used 4 versions generated by the models, resulting in a total number of 240 test items (48×5).

4.2 Experiment Design

The 240 items were divided into 16 groups of 15 items each through a pseudo-randomisation process. This process ensured, to the greatest extent possible, that two versions of the same test item did not appear in the same group.

At the beginning of the experiment, we explained the goal of the experiment to the participants in broad terms, and we clarified what we expected them to do. The full instruction can be found in Appendix B.

Since our participants were not linguistic experts, we opted for simpler terminology, using "expression" instead of "referring expression" and "paragraph" instead of "text" or "discourse". A slight drawback of this formulation is that it does not forbid the use of phrases that are not REs or even noun phrases when suggesting rewritings. In fact, however, since the non-REs were very infrequent in the outcomes and we felt that this should not affect the testing of our hypotheses, we processed all rewritings in the same way (see Section 5.1).

Each item contains several questions and tasks that participants need to accomplish, an example of which is shown in Figure 1. It starts by showing participants the whole item's text, which is followed by the following questions/tasks.[5]

Familiarity. The WEBNLG dataset contains data units that are composed of RDF triples, each extracted from DBPedia. The accompanying texts for these data units are sequences of one or more sentences that verbalise the information in the RDF triples (Gardent et al., 2017b). These texts revolve around a central entity. We call this entity the *major entity* in the discourse. The first question in the experiment is a Yes-No question, asking whether the participant is familiar with the major entity. In the question, we referred to the major entity using its proper name, which is obtained by replacing underscores in its identifier (hereafter, PROPER NAME).

Referential Success. The second question asks whether the RE in question is successful in identifying the referent. This question was asked only for REs that differ from their PROPER NAME. For instance, in the example in Table 1, the RE "*the school*" is one such case. This expression differs from the proper name format of the identifier. The identifier, in this case, is "AWH_Engineering_College", and its PROPER NAME is "AWH Engineering College". We highlighted the RE and asked whether the expression "*the school*" refers to "AWH Engineering College".

Moreover, in our pilot study, we found that participants were sometimes unsure whether an RE was successful or not. Thus, we added a "Maybe" option. This option could also provide us with insights about the REs that are more difficult to resolve.

It is worth noting that WEBNLG contains a few errors. For example, it occasionally marks "American" as referring to the United States. We manually corrected these errors while preparing the experiment materials.

[3]The data from Same et al. (2022) is available at https://github.com/a-quei/neuralreg-re-evaluation.

[4]The full results of these models from Same et al. (2022) can be found in Appendix A

[5]Henceforth, the term "text" will refer to any one of the short paragraphs that we presented to participants and that can be understood by itself.

AWH Engineering College is in Kuttikkattoor, India in the state of Kerala. The school has 250 employees and Kerala is ruled by Kochi. The Ganges River is also found in India.

Question 1: The above text is a short introduction to AWH Engineering College. Were you aware of the existence of AWH Engineering College before this experiment? [Yes, No]

Question 2: Regarding the above paragraph, please answer the question(s) below.
1. Does the expression "The school" (highlighted) in the text below refer to "AWH Engineering Colleg"? [Yes, Maybe, No]
AWH Engineering College is in Kuttikkattoor, India in the state of Kerala. ***The school*** *has 250 employees and Kerala is ruled by Kochi. The Ganges River is also found in India.*
2. (...)

Question 3: We have highlighted some of the expressions in the paragraph. Please focus on these expressions when answering the questions below.
AWH Engineering College *is in* ***Kuttikkattoor***, ***India*** *in the state of* ***Kerala***. ***The school*** *has 250 employees and* ***Kerala*** *is ruled by* ***Kochi***. ***The Ganges River*** *is also found in* ***India***.
Focusing on these expressions, to what extent do you think the following statements are true?
- This paragraph is clear. [1-7]
- This paragraph is grammatical. [1-7]
- This paragraph is coherent. [1-7]
Below, we have listed these expressions in the paragraph. Could you suggest a better alternative for each expression to enhance the paragraph's coherence, grammatical correctness, and clarity? For the one that you think is optimal, you can simply copy and paste the expression in the paragraph into the text box. (For your convenience, we show you the highlighted paragraph again with each expression numbered).
(1) AWH Engineering College *is in* ***(2) Kuttikkattoor***, ***(3) India*** *in the state of* ***(4) Kerala***. ***(5) The school*** *has 250 employees and* ***(6) Kerala*** *is ruled by* ***(7) Kochi***. ***(8) The Ganges River*** *is also found in* ***(9) India***.
- (1) AWH Engineering College:
- (2) (...)

Apart from these expressions, do you have any other comments or suggestions? (Optional)

Figure 1: An example item in our experiment.

Rating. Given our aim to perform a quantitative evaluation of REG models and our curiosity about the impact of the meta-linguistic tasks on the overall scores, we asked participants (in the first part of the third question, see Figure 1) to rate the text in the same manner as described in Same et al. (2022). Concretely, participants were asked to answer whether they agreed with the following three statements on a 7-point Likert-scale, where 1 means "strongly disagree", 4 means "I don't know", and 7 means "strongly agree": (1) Clarity: This paragraph is clear; (2) Grammaticality: This paragraph is grammatical; and (3) Coherence: This paragraph is coherent.[6] Since factors other than the quality of REs can also influence the overall quality of a text, we asked participants to rate the texts while focusing on the REs highlighted in the discourse.

Rewriting. As discussed in Section 3, finally, we asked participants to suggest better rewritings in the second part of question 3 (see the example in Figure 1). Participants were instructed to "*suggest a better alternative for each expression to enhance the paragraph's coherence*". To make sure that no RE was overlooked or skipped, we made it mandatory for them to write a suggestion for each RE. We mentioned that if they found an RE to be optimal, they could simply copy and paste the original RE into the designated slot.

It is worth noting that we placed the rewriting task after the rating task because we expected that

[6]Recently, it has been pointed out that terms like "coherence" or "fluency" are vague for participants who do not have linguistic background, causing more variations in responses (Howcroft et al., 2020). In this work, we kept using "coherence" to ensure our experimental setting was identical to Same et al. (2022).

completing the rating first would lead the participants to read the entire text before beginning to optimise its REs. We combined these two tasks into a single question with the expectation that participants might re-evaluate their scores following the rewriting task.

Additional Comments. During the pilot study, we observed that some participants criticised non-referential aspects of the text (i.e. aspects that did not concern the REs in the text). Therefore, in the main experiment, we allowed participants to provide comments on aspects other than just REs.

4.3 Participants and Procedure

We constructed the experiment materials using Qualtrics and carried out the experiment on Prolific. Each set of items was completed by 8 participants. We restricted participants to native speakers of English located in the United Kingdom or the United States. Based on the pilot study, we expected that participants would complete the experiment, which consisted of 15 items, within 30 minutes. Therefore, we paid Proflic 5 GBP for each participant. We rejected participants if they (1) apparently misunderstood any of our tasks or (2) wrote nonsensical responses.

Besides the demographic information available on prolific, we did not collect any additional personal information during the experiment.

We obtained responses from 128 participants (16×8), with an average age of 38. Of these, 75 identified as female, and 51 identified as male. The average duration of the experiment was 41 minutes, which was higher than what we expected. This was because several participants had strong views on the use of RE. They tried to optimise every RE and left very long comments, which made them spend more than an hour on our experiment.

5 Results

In this section, we introduce the data we obtained from the experiment, explain how we post-processed the data, and report the results.

5.1 The Dataset

There are a total of 1325 REs in the 240 test items. Out of these, 469 REs are different from their PROPER NAMEs. We asked about the referential success of these REs and received 3752 responses (469×8).

For these 1325 REs, we obtained 10600 participant-written REs from our participants. These REs may contain typos or formatting issues; for example, some participants wrote short comments in the text boxes intended for writing suggested REs. Therefore, we manually corrected every participant-written RE and annotated whether the RE was a rewriting or a copy of the original RE. Ultimately, we obtained 2832 rewritings.

It is worth noting that, during annotation, we found that some participants commented on certain REs that "*it is impossible to infer which referent this RE refer to*" or "*this RE is redundant*". We annotated the former case as "unresolvable" and the latter case as "redundant". Out of 10600 participant-written REs, we identified 53 "unresolvable" cases for 48 REs (an RE can be marked "unresolvable" by multiple participants) and 39 "redundant" cases for 17 REs. We treated these cases as rewritings. The data is available at: https://github.com/a-quei/reg-rewriting.

5.2 Main Results

Referential Success. Table 2 presents the results for the answers to questions that ask about the referential success of REs. It includes the proportion of each response type ('Yes', 'Maybe', and 'No') for each model. Additionally, we computed the Success Rate (SR), which is defined as the number of 'Yes' responses a model received to the product of the number of REs and the number of participants. SR considers PROPER NAMEs (see Section 4) as referentially successful REs. These numbers for Seen and Unseen portions are also reported. The raw count of each answer can be found in Appendix C.

RULE has the highest SR among all other models, including Human. This is because it used PROPER NAMEs in the majority of cases and, for the rest, chose only pronouns. By never using descriptions, the REs it generated were easier to resolve. The SR of ML is on par with Human. This model works remarkably well on unseen data.[7] None of the REs it generated for this data were ever marked as being definitely unsuccessful. On the seen data, the REs generated by ML and marked as "maybe" were often pronouns with slight referential ambiguity. Compared to RULE and ML, the two neural models were more likely to produce ambiguous REs. Nonetheless, the REs they generated for seen data

[7] Recall that some models work better on unseen data, primarily because this data is simpler than seen data. See Section 2 for more details.

	All				Seen				Unseen			
Model	SR	Yes	Maybe	No	SR	Yes	Maybe	No	SR	Yes	Maybe	No
RULE	99.39	90.97	6.25	2.68	99.39	90.28	5.56	4.17	99.39	91.67	6.94	1.39
ML	94.14	77.01	17.63	5.36	91.08	75.94	18.40	5.66	99.90	95.83	4.17	0.00
NREG-1	80.80	64.67	25.87	9.46	93.18	83.75	11.67	4.58	66.29	51.04	36.01	12.95
NREG-2	81.51	64.23	25.82	9.95	90.21	78.12	15.23	6.64	71.31	52.05	35.10	12.84
Human	94.62	87.50	9.32	3.18	93.79	85.45	10.04	4.51	95.59	89.86	8.49	1.65

Table 2: The proportion of each response to the questions concerning the referential success of REs. SR stands for successful rate of REs.

are equally as successful as Human, but those for unseen data are dramatically worse.

Ratings. Table 3 reports the participants' ratings. Compared to Same et al. (2022) (cf. Table 5 in Appendix A), on the same set of test samples, scores from our experiment are significantly lower (in terms of clarity, grammaticality, and coherence using a Mann-Whitney Test; $p < .001$), which confirms our hypothesis $\mathcal{H}_2$.

Unlike Same et al. (2022), Human (the original texts) achieves significantly better performance than all the other models in terms of all three criteria (using Wilcoxon's signed-rank test with Bonferroni correction). The experimental models seem to be still indistinguishable from each other except NREG-2, which has the lowest clarity score.

Zooming in on the seen data, the two neural models perform equally well to Human while non-neural models (especially RULE) receive lower grammaticality and coherence scores. On unseen data, the situation is the other way around. Since unseen data has lower complexity, RULE and ML can produce clear and coherent REs, but the grammaticality of these REs is still a problem. Meanwhile, neural models are significantly worse than other models and Human. In short, we observe significant differences in ratings, confirming $\mathcal{H}_1$.

Moreover, these findings are consistent with the analysis of the referential success of REs, revealing that non-neural models may struggle with processing complex situations, while neural models find it difficult to handle entities that they have never seen. Such phenomena were not supported by the outcomes of Same et al. (2022) (cf. Appendix A).

Rewriting. We quantify the results of the rewriting task by computing the rewriting rates (RR), defined as the proportion of rewritings over all REs generated by each model. The results of RR are also depicted in Table 3 (raw counts of re-writings can be found in Appendix C), from which we observe a similar trend to the previous results. Models that receive higher scores generally have lower RR. Neural models have low RR on seen data and high RR on unseen data, while non-neural models have similar RR scores on both seen and unseen data.

5.3 Relations between Ratings and Responses to the Tasks

To test $\mathcal{H}_3$, we first tested how the referential success is correlated with clarity scores. To this end, we examine whether there is a significant positive correlation between clarity and the number of successful REs in a text (i.e. the sum of 'YES' responses from our experiment and the number of PROPER NAMEs) and the number of 'YES' responses alone.[8] We did linear regression tests (i.e. using each measure above to predict the clarity score) and reported the p-value as well as the slope (β) and the effect size computed using the coefficient of determination (R^2). The results show that both the number of 'YES' responses ($\beta = .12, R^2 = .0099, p < .001$) and the number of successful REs ($\beta = .057, R^2 = .0048, p = .002$) positively correlate with clarity. This confirms $\mathcal{H}_3$.

Regarding the relation between the ratings and participants' rewritings of each RE ($\mathcal{H}_4$), we used a linear regression test to assess the correlation of the number of rewritings on each rating. We identified significant negative impacts on clarity ($\beta = -.34, R^2 = .083, p < .001$), grammaticality ($\beta = -.36, R^2 = .081, p < .001$), and coherence ($\beta = -.37, R^2 = .089, p < .001$). These findings suggest an affirmative answer to $\mathcal{H}_4$.

As for the last hypothesis, $\mathcal{H}_5$, we compared scores from items where participants were familiar with the major entity to those where they were not. Mann Whitney U tests found no significant difference in clarity, grammaticality, coherence scores, or

[8] Recall that all REs in a text may be successful, but since they are all PROPER NAMEs, participants were never asked about their success in our experiment.

	All				Seen				Unseen			
Model	C	G	Co	RR	C	G	Co	RR	C	G	Co	RR
RULE	4.78^{B}	4.12^{B}	4.62^{B}	28.44	4.61^{A}	3.84^{B}	4.33^{B}	29.81	4.95^{A}	4.40^{B}	4.92^{A}	26.84
ML	4.79^{B}	4.34^{B}	4.63^{B}	26.51	4.61^{A}	$4.28^{A,B}$	4.45^{B}	26.05	4.98^{A}	4.40^{B}	4.81^{A}	27.05
NREG-1	$4.49^{B,C}$	4.18^{B}	4.36^{B}	27.31	5.01^{A}	4.51^{A}	$4.81^{A,B}$	20.37	3.98^{B}	3.86^{B}	3.91^{B}	35.45
NREG-2	4.40^{C}	4.15^{B}	4.33^{B}	30.90	4.79^{A}	4.43^{A}	$4.70^{A,B}$	24.30	4.01^{B}	3.88^{B}	3.96^{B}	38.63
Human	5.17^{A}	4.85^{A}	5.10^{A}	20.42	5.05^{A}	4.63^{A}	4.97^{A}	21.33	5.30^{A}	5.06^{A}	5.22^{A}	19.36

Table 3: The rating results and re-writing rates (RR). 'C', 'G' and 'Co' stand for Clarity, Grammaticality, and Coherence, respectively. Rankings are determined by significance testing ($p < 0.01$; using Wilcoxon's signed-rank test with Bonferroni correction). Per column, results that have *no* superscript letters in common are significantly different from each other. Note that the lower the RR, the better.

the number of rewritings. Therefore, we accepted the null hypothesis, suggesting that familiarity did not play a role in relation to this corpus.

6 Further Observations

Additionally, we made the following observations while annotating the data.

6.1 Additional Comments

During the experiment, we received a few comments, which were supposed to focus on issues in the test items other than the contents of REs. These comments can be categorised into 4 types: (1) Potential ambiguities in the given text; (2) Ethical issues, for instance, "*Though the discourse is generally coherent, the pronouns are better neutralised (e.g., using 'they' instead of 's/he')*"; (3) Inaccurate/inappropriate phrases other than REs; (4) Overall quality: quite a few comments suggest that the text as a whole is of low quality, for example, "*the language is unprofessional*" or "*the discourse structure of the paragraph is bad*".

6.2 Types of Rewritings

We can use rewritings as a proxy to analyse the incorrectness and inappropriateness of REs. Thus, the rewritings can be roughly divided into two categories, as follows.

6.2.1 Correcting Errors in REs

Some rewritings are about correcting errors in REs, including the following error types: (1) typo or grammatical error ("*a admiral*" → "*an admiral*"); (2) degeneration ("*the Koc Koc*" → "*The Koc*"); (3) definiteness ("*AWH College*" → "*the AWH College*"); (4) possessive ("*Alan Frew*" → "*Alan Frew's*"); (5) unknown referent: participants sometimes pointed out that given the RE and its context, it was not possible to infer which referent it refers to. This happened when there was serious referential ambiguity or the non-pronominal form of the referent had never appeared in the previous discourse; (6) incorrect referent: some rewritings changed the RE to refer to a completely different referent. It happened when (a) the context of the RE suggests that the RE at this position should refer to a different referent, (b) the RE is a pronoun, and there is a mismatch in its surface form, and (c) the RE contradicts the common knowledge of the participant (e.g., we observed that multiple participants rewrote "*Elizabeth II*" to "*Charles III*").

6.2.2 Optimising REs

Other rewritings aim at optimising the content of REs to make the whole discourse clearer and more coherent, including the following kinds: (1) referential form ("*Alan Frew*" → "*he*"); (2) punctuation ("*the icebreaker Aleksey Chirikov*" → "*the icebreaker, Aleksey Chirikov,*"); (3) paraphrasing: some REs are paraphrased to be more readable ("*the defender (football)*" → "*the football defender*"); (4) elaboration/simplification: some REs were considered to be over-specificified or under-specified (Chen and van Deemter, 2023) and, thus, were simplified or elaborated in the rewritings; (5) non-RE: since our experiment did not limit participants to filling in only REs, rewritings could also be something other than referring expressions; (6) style (e.g., politeness: "*Alan Shepard*" → "*Mr. Alan Shepard*"); (7) ethical issues ("*he*" → "*they*"). In a follow-up study, we plan to conduct a qualitative analysis of the rewritings and a detailed annotation of the rewriting types observed.

6.3 The Context of an RE

As discussed, REG is highly context-dependent. Nonetheless, almost all computational REG mod-

els so far consider merely textual context.[9] In this study, we observed multiple other kinds of contexts that also play important roles in humans' use of RE. First, the optimality of REs in a discourse depends on each other. For example, participants sometimes rewrote REs to avoid duplication. However, the current setting of End2End REG models does not allow dependent production of REs. Second, the style of text contributes greatly to how elaborate the REs should be. In our case, since our data was in Wikipedia style, many participants thought the REs should be as elaborate as they could. As a consequence, for instance, they viewed "*Essex County*" as an unsuccessful RE (although it is a distinguishing proper name) and rewrote it to "*Essex County, New York*". Last, we observed quite a lot of cases where the background knowledge is influential. For example, participants from the United Kingdom often rewrote "*Elizabeth II*" to "*Charles III*" while those from the United States often rewrote "*Donald Trump*" to "*Joe Biden*" when referring to the leader of the country.

7 Conclusion

Focusing on some surprising results from Same et al. (2022), namely that REG models are indistinguishable from each other and from the REs in WEBNLG in a rating-based human evaluation, this paper has introduced a new type of intrinsic task-based REG evaluation. In parallel with rating, we designed two tasks: one asks participants about the referential success of each RE, and the other asks participants to suggest a rewriting for each RE if possible. In this way, we had a better understanding of REG models' performance from different perspectives. Meanwhile, we confirmed that accomplishing these meta-level tasks helped participants rate in a more reliable and discriminable way.

This comprehensive evaluation suggests that, on WEBNLG, the machine learning based REG model performed best of all the models we tested. Compared to the rule-based and neural models, it had a remarkably high rate of producing referentially successful REs. Additionally, it received the best clarity, grammaticality and coherence scores, and its generated content was rewritten the least frequently.

Same et al. (2022) highlighted the usefulness of non-neural models because the classical rating-based human evaluation suggested that neural and non-neural models perform at comparable levels, at least in the area of referring expression generation. The results from our new intrinsic task-based evaluation reinforce earlier conclusions because they suggest that, in fact, non-neural models outperform neural models in this area.

We think our design of the intrinsic task-based human evaluation protocol can serve as a reference for the evaluation of other NLG tasks that are complex enough so that simple evaluation cannot offer all the answers. We also think that the data from the experiment can help linguists understand the use of reference better and help computer scientists build better REG models.

In future, on the one hand, we plan to conduct more in-depth analyses of the issues we discussed in Section 6, including qualitative and quantitative analyses of the rewritings and the factors that could affect the use of RE. On the other hand, our results revealed that neural models are good at processing seen data but not unseen data. This problem might be addressed by the most recent Large Language Models (LLMs).[10] For this reason, in future research, we plan to apply mainstream LLMs to REG and assess these models using the protocol proposed and investigated in this paper.

Ethical Considerations

Two ethical considerations need to be noted here: First, we recruited participants on Prolific and only used their demographic information that is publicly available on Prolific. Our experiment does not collect any personal information. Second, a few participants reported that referring expressions in WEBNLG or those produced by REG models might be gender-biased.

Limitations

In this paper, we used the term "neural model" to refer to the NeuralREG models that we tested in this work and used the "state-of-the-art" to refer to the state-of-the-art when Same et al. (2023) was carried out. As explained, our specific conclusions about neural models may not generalise to the most recent pre-trained Large Langauge Models. Second, this work considered only a simple dataset,

[9] Exceptions include, for example, Cao and Cheung (2019), who considered the knowledge about the major entity.

[10] We did not test LLMs as (1) our focus in this paper is not on seeking the best-performing REG models, and (2) this approach allows us to reuse materials from Same et al. (2022), ensuring a fair comparison.

namely, WEBNLG. It has been argued that WEBNLG is flawed as a tool for REG evaluation (Chen et al., 2023) and the choice of the corpus would highly influence the evaluation results (Same et al., 2023). It is worth noting that, in this paper, we have only challenged the evaluation protocol used in previous studies; our findings do not focus on the choice of evaluation corpus. Finally, in our experiment, we asked participants to only rewrite REs (rather than rewriting the entire text). This might somewhat decrease the ecological validity of the experiment, as humans normally do not produce REs given linguistic contexts that have already been realised.

Acknowledgments

We are grateful for the comments from reviewers. Guanyi Chen is supported by the start-up funds of Central China Normal University (31101232053).

References

Anja Belz and Sebastian Varges. 2007. Generation of repeated references to discourse entities. In *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)*, pages 9–16, Saarbrücken, Germany. DFKI GmbH.

Luisa Bentivogli, Mauro Cettolo, Marcello Federico, and Christian Federmann. 2018. Machine translation human evaluation: an investigation of evaluation based on post-editing and its relation with direct assessment. In *Proceedings of the 15th International Conference on Spoken Language Translation*, pages 62–69, Brussels. International Conference on Spoken Language Translation.

Meng Cao and Jackie Chi Kit Cheung. 2019. Referring expression generation using entity profiles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3163–3172, Hong Kong, China. Association for Computational Linguistics.

Thiago Castro Ferreira, Diego Moussallem, Ákos Kádár, Sander Wubben, and Emiel Krahmer. 2018a. NeuralREG: An end-to-end approach to referring expression generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1969, Melbourne, Australia. Association for Computational Linguistics.

Thiago Castro Ferreira, Diego Moussallem, Emiel Krahmer, and Sander Wubben. 2018b. Enriching the WebNLG corpus. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 171–176, Tilburg University, The Netherlands. Association for Computational Linguistics.

Guanyi Chen, Fahime Same, and Kees van Deemter. 2023. Neural referential form selection: Generalisability and interpretability. *Computer Speech & Language*, 79:101466.

Guanyi Chen and Kees van Deemter. 2023. Varieties of specification: Redefining over-and underspecification. *Journal of Pragmatics*, 216:21–42.

Rossana Cunha, Thiago Castro Ferreira, Adriana Pagano, and Fabio Alves. 2020. Referring to what you know and do not know: Making referring expression generation models generalize to unseen entities. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2261–2272, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017a. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188, Vancouver, Canada. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017b. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Int. Res.*, 61(1):65–170.

Navita Goyal, Ani Nenkova, and Hal Daumé III. 2023. Factual or contextual? disentangling error types in entity description generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8322–8340, Toronto, Canada. 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.

Roman Kutlak, Kees Van Deemter, and Chris Mellish. 2011. Audience design in the generation of references to famous people. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

Ehud Reiter. 2011. Task-based evaluation of nlg systems: Control vs real-world context. In *Proceedings*

of the UCNLG+ Eval: Language Generation and Evaluation Workshop, pages 28–32.

Ehud Reiter. 2017. A commercial perspective on reference. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 134–138, Santiago de Compostela, Spain. Association for Computational Linguistics.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Studies in Natural Language Processing. Cambridge University Press.

Fahime Same, Guanyi Chen, and Kees Van Deemter. 2022. Non-neural models matter: a re-evaluation of neural referring expression generation systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5554–5567, Dublin, Ireland. Association for Computational Linguistics.

Fahime Same, Guanyi Chen, and Kees van Deemter. 2023. Models of reference production: How do they withstand the test of time? In *Proceedings of the 16th International Natural Language Generation Conference*, pages 93–105, Prague, Czechia. Association for Computational Linguistics.

Somayajulu Sripada, Ehud Reiter, and Lezan Hawizy. 2005. Evaluation of an NLG system using post-edit data: Lessons learnt. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*, Aberdeen, Scotland. Association for Computational Linguistics.

Ieva Staliūnaitė, Hannah Rohde, Bonnie Webber, and Annie Louis. 2018. Getting to "hearer-old": Charting referring expressions across time. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4350–4359, Brussels, Belgium. Association for Computational Linguistics.

Craig Thomson, Ehud Reiter, and Barkavi Sundararajan. 2023. Evaluating factual accuracy in complex data-to-text. *Computer Speech and Language*, 80:101482.

Kees van Deemter. 2016. *Computational models of referring: a study in cognitive science*. MIT Press.

Weischedel, Ralph, Palmer, Martha, Marcus, Mitchell, Hovy, Eduard, Pradhan, Sameer, Ramshaw, Lance, Xue, Nianwen, Taylor, Ann, Kaufman, Jeff, Franchini, Michelle, El-Bachouti, Mohammed, Belvin, Robert, and Houston, Ann. 2013. OntoNotes release 5.0 LDC2013t19.

A Results in Same et al. (2022)

Table 5 shows the results for the models we examined in this work from Same et al. (2022). Since, in this study, we tested fewer models, the factors of the Bonferroni correction would be different. Therefore, we re-did the significant testing and reported the results in Table 5. Additionally, we found a small error in the results in Same et al. (2022), which is also corrected in Table 5.

B Instruction of Our Experiment

We explained the general goal to the participants and clarified what we expected them to do:

> *In this experiment, you will see 15 short paragraphs, each containing 2-3 sentences. We are particularly interested in the use of some of the expressions within these paragraphs. Accordingly, we will ask you to answer several questions about these expressions in the paragraphs. Given that language use can often be imperfect, the final question will ask you to suggest better alternatives for each expression in order to improve the clarity, grammaticality, and coherence of the paragraph.*

C Raw Numbers from Our Experiment

Table 6 reports the raw number of answers to the questions about the referential success of REs. In addition to the findings in Section 5, the numbers here show that ML rarely produced REs that are not proper names when processing unseen data. Table 4 charts the number of rewritings.

Model	All	Seen	Unseen
RULE	603	341	262
ML	562	298	264
NREG-1	579	233	346
NREG-2	655	278	377
Human	433	244	189

Table 4: The number of rewritings.

	All			Seen			Unseen		
Model	Clarity	Grammar	Coherence	Clarity	Grammar	Coherence	Clarity	Grammar	Coherence
RULE	5.71^{A}	5.73^{A}	5.76^{A}	5.68^{A}	5.62^{A}	5.73^{A}	5.75^{A}	5.83^{A}	5.79^{A}
ML	5.67^{A}	5.63^{A}	5.78^{A}	5.62^{A}	5.63^{A}	5.73^{A}	5.72^{A}	$5.62^{A,B}$	5.82^{A}
NREG-1	5.68^{A}	5.62^{A}	5.65^{A}	5.76^{A}	5.64^{A}	5.71^{A}	5.59^{A}	$5.60^{A,B}$	5.58^{A}
NREG-2	5.66^{A}	5.56^{A}	5.68^{A}	5.65^{A}	5.68^{A}	5.69^{A}	5.66^{A}	5.43^{B}	5.67^{A}
Human	5.82^{A}	5.69^{A}	5.81^{A}	5.83^{A}	5.69^{A}	5.77^{A}	5.80^{A}	$5.70^{A,B}$	5.84^{A}

Table 5: Human Evaluation Results from Same et al. (2022) on the WEBNLG corpus. Rankings are determined by significance testing ($p < 0.01$; using Wilcoxon's signed-rank test with Bonferroni correction). Per column, results that have *no* superscript letters in common are significantly different from each other.

	All			Seen			Unseen		
Model	Yes	Maybe	No	Yes	Maybe	No	Yes	Maybe	No
RULE	131	9	4	65	4	3	66	5	1
ML	345	79	24	322	78	24	23	1	0
NREG-1	745	298	109	402	56	22	343	242	87
NREG-2	704	283	109	400	78	34	304	205	75
Human	798	85	29	417	49	22	381	36	7

Table 6: The count of each answer to the questions concerning the referential success of REs.