Two Reproductions of a Human-Assessed Comparative Evaluation of a Semantic Error Detection System

Rudali Huidrom
ADAPT/DCU, Dublin, Ireland
rudali.huidrom@adaptcentre.ie

Ondřej Dušek and **Zdeněk Kasner**
Charles University, Prague, Czechia
{odusek,kasner}@ufal.mff.cuni.cz

Thiago Castro Ferreira
Federal University of Minas Gerais, Brazil
Thiagocf05@ufmg.br

Anya Belz
ADAPT/DCU and University of Aberdeen
anya.belz@adaptcentre.ie

Abstract

In this paper, we present the results of two reproduction studies[1] for the human evaluation originally reported by Dušek and Kasner (2020) in which the authors comparatively evaluated outputs produced by a semantic error detection system for data-to-text generation against reference outputs. In the first reproduction, the original evaluators repeat the evaluation, in a test of the *repeatability* of the original evaluation. In the second study, two new evaluators carry out the evaluation task, in a test of the *reproducibility* of the original evaluation under otherwise identical conditions.[2] We describe our approach to reproduction, and present and analyse results, finding different degrees of reproducibility depending on result type, data and labelling task. Our resources are available and open-sourced[3].

1 Introduction

Reproduction studies are garnering growing interest in natural language processing (NLP), most recently as the subject of shared tasks (Branco et al., 2020; Belz et al., 2021). The importance of ensuring good levels of reproducibility in NLP work is increasingly recognised, and approaches to defining and assessing reproducibility are emerging (Cohen et al., 2018; Belz et al., 2022). With this paper, we add to the growing body of reproduction studies by tackling a particularly hard case for reproducibility assessment, namely error analysis that involves identifying which of two disagreeing systems is making error(s), and further classifying the types of errors being made.

We perform two reproductions, one involving the same evaluators as in the original study, one involving new evaluators. The former can be seen as a test of the repeatability of the original study, where nothing is changed except the point in time, and the latter as a test of its reproducibility where the reproduction differs from the original study in some specfied respect(s), here the evaluator cohort.[4]

Below we start by describing the original study and outlining our approach to reproduction. Next we describe our two reproductions and present an analysis which examines three types of results from the evaluations, applying different tools for measuring similarity in each case. We finish with a discussion of the reasons behind and possible mitigation strategies for what is, on the face of it, a mostly poor set of reproducibility results.

2 Original Evaluation

2.1 Semantic error detection method

Dušek and Kasner (2020) presented an automatic method for semantic error detection (SED) in data-to-text generation (see Figure 1 for example data/text pairs) based on textual entailment checking. The basic idea is to trivially (and automatically) map each triple in the input meaning representation (top part of each example in Figure 1) to a text representation using simple generation templates, and then to check whether input and output entail each other. If the input does not entail the output, a hallucination error is diagnosed (some content in the output is not present in the input); if the output does not entail the input, it is taken to

[1] Carried out as part of the ReproGen 2022 shared task.

[2] With the proviso that instructions had to be created for the reproductions.

[3] https://github.com/RHuidrom/reprogen22_dusek_and_kasner_2020.git

[4] See also Section 4 re new instructions.

Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges, pages 52 - 61

MR: Atlantic City, New Jersey | country | United States
United States | capital | Washington, D.C.

NLG system output: atlantic city, new jersey comes from the united states where the capital is washington, d.c.

SED label: Reference label (derived from human rating): *not OK*; NLI-SED label: *OK*

Error analysis annotation: *other* (both system and reference are incorrect), *bad sentence*

MR: FC Dinamo Batumi | manager | Levan Khomeriki
Aleksandre Guruli | club | FC Dinamo Batumi

NLG system output: fc dinamo batumi was at levan khomeriki and manages aleksandre guruli.

SED label: Reference label (derived from ruman rating): *not OK*; NLI-SED: *OK*

Error analysis annotation: *reference correct*, *unjustified OK*

Figure 1: Two examples each consisting of a meaning representation (MR); an NLG system output (from WebNLG 2017); two SED labels (the reference error label derived from the WebNLG 2017 human ratings, and the output from the NLI-SED system); and correctness and error label annotations as produced in one of our reproductions.

mean an omission error (some content in the input is not present in the output). If input and output do entail each other, then the output is taken to be error-free.

For the entailment checking, the method used a pretrained RoBERTa model (Liu et al., 2019) from the Transformers library (Wolf et al., 2020) finetuned on the MultiNLI dataset (Williams et al., 2018). The model (referred to as the NLI-SED system below) produces probability estimates for the three possible outputs: contradiction, neutral and entailment. To pass an entailment check, the entailment probability simply has to be higher than the neutral and contradiction probabilities.

When checking whether the output entails the input, Dušek and Kasner paired the simple text representation of each triple with all of the output text and performed the entailment check on each pair individually. When checking whether the input entails the output, the simple text representations of all input triples were concatenated and paired with the output text in a single entailment check.

Ultimately, the output from Dušek and Kasner's NLI-SED system is one of the following: *OK, omission, hallucination, hallucination+omission.*

2.2 Manual evaluation of the SED method

The original study that is the subject of reproduction in this paper is a manual evaluation in which Dušek and Kasner compared the SED labels obtained from their NLI-SED system for data from the E2E (Dušek et al., 2020) and the WebNLG (Gardent et al., 2017) shared tasks with reference labels. They performed an error analysis on a sample of 100 cases where NLI-SED system generated label and reference label disagreed. In each case, they decided which was right and which was wrong, and additionally selected labels indicating the likely source(s) of any error(s), from among six different error labels for E2E, and five for WebNLG (labels as described for each dataset below).[5] Finally, in each case, the authors also provided unstructured notes which explain their annotations.

E2E		
Counts of	Slot-error Script	NLI-SED
OK	33	54
omission	42	32
hallucination	17	7
omiss+halluc	8	7
WebNLG		
Counts of	Human Ratings	NLI-SED
OK	45	54
not OK	55	46

Table 1: Counts for different SED labels as per the reference labels (produced by the slot-error script in the case of E2E, and derived from human ratings in the case of WebNLG), and the NLI-SED system.

2.2.1 E2E

For E2E, reference labels (*OK, omission, hallucination, hallucination+omission*) were available from the E2E shared task where they were generated by the organisers with what they termed a slot-error script based on regular expression matching, with patterns informed by a subset of the E2E development set.

In the sample Dušek and Kasner annotated in their error analysis, the counts for reference labels produced by the slot-error script and for the NLI-SED system generated labels look as shown in Table 1. In addition, there was partial agreement between the reference labels from the slot-error script

[5]The error classes and raw counts from the annotations we use in this paper were not reported in the original publication, but were instead mapped to less fine-grained findings.

Counts of	E2E					Counts of	WebNLG				
	Dušek & Kasner 2020	Repeat. Test (A1+A2)	*CV**	Reprod. Test (A3+A4)	*CV**		Dušek & Kasner 2020	Repeat. Test (A1+A2)	*CV**	Reprod. Test (A3+A4)	*CV**
ref correct	34	36	5.697	41	18.611	ref correct	51	38	29.126	59	14.502
SED correct	45	48	6.432	44	2.240	SED correct	42	40	4.863	35	18.127
other	18	16	11.730	15	18.127	other	7	15	72.510	6	15.339
[eatType]	5	6	18.127	6	18.127	[bias-templ]	22	16	31.484	5	125.549
[priceRange]	30	33	9.495	28	6.876	[val-format]	7	3	79.760	10	35.188
[famFriend]	10	13	26.019	8	22.156	[bad-sent]	14	27	63.225	10	33.234
[f-halluc]	8	5	46.016	22	93.054	[unj-OK]	8	25	102.722	28	110.778
[f+omiss]	16	11	36.926	24	39.880	[unj-notOK]	15	19	23.460	12	22.156
[f+halluc]	17	20	16.168	8	71.784						

Table 2: QRA assessment of correctness and error label counts (type *i* results), on the ***combined*** annotations (in Repeatability Test, half randomly taken from each original annotator; in Reproducibility Test, half randomly taken from each of the new annotators).

and NLI-SED system in 12 cases, where both detected an omission (and one additionally detected a hallucination). There was no partial agreement on hallucinations.

In Dušek and Kasner's annotations, the script-generated labels were deemed to be correct (and the NLI-SED system's prediction wrong) in 34 out of 100 cases, and the NLI-SED system's predictions were deemed correct (and the script wrong) in 45 cases. In 18 cases, either both the script-generated labels and the NLI-SED system's prediction were wrong or the evaluators were unable to decide. These numbers are also included in the upper part of the first Dusek & Kasner column in Table 2.[6]

The six error class types for the E2E error annotations were as listed below. Note that the descriptions and definitions given here were created as part of our reproductions. The implications of creating new instructions for a reproduction are discussed in Section 5.

Each error class represents a different possible source of an error made by Dušek and Kasner's NLI-SED system or the slot-error script, and as many error classes were selected as applied in each case, in some cases none were selected (frequencies are shown in the lower part of the first Dusek & Kasner column in Table 2). These error classes tend to apply predominantly to either the NLI-SED system or the slot-error script, indicated by underlines below. The short labels in square brackets are used to refer to each class in the results tables below.

1. **Error related to *eatType=restaurant* slot value [eatType]:** The incorrect SED label (produced either by the NLI-SED system or the slot error script) is likely caused by something involving the slot/value pair eatType=restaurant, e.g. not detecting a hallucination when the eatType slot is not in the input, but the output mentions a restaurant.
2. **Error related to *priceRange* slot [priceRange]**: The incorrect SED label (produced by either the NLI-SED system or the slot error script) makes an error related to the priceRange slot, e.g. incorrectly identifying a hallucination in the priceRange slot, when the price range information has in fact been correctly verbalised.
3. **Error related to *familyFriendly attribute* [famFriend]**: The incorrect SED label (produced by either the NLI-SED system or the slot error script) makes an error related to the familyFriendly slot, e.g. incorrectly identifying an omission when the information has in fact been correctly verbalised.
4. **Other false negative hallucination ('off-topic blabber') [f-halluc]**: The incorrect SED label (produced by either the NLI-SED system or the slot error script) fails to detect a hallucination (unrelated to E2E slots) present in the verbalisation.
5. **Other false positive omission ('unjustified omission') [f+omiss]:** The incorrect SED label (produced by either the NLI-SED system or the slot error script) wrongly detects an omission in the verbalisation.
6. **Other false positive hallucination ('unjustified hallucination') [f+halluc]:** The in-

[6] Numbers don't add up to 100 because of missing annotations.

correct SED label (produced by either the NLI-SED system or the slot error script) wrongly detects a hallucination in the verbalisation.

2.2.2 WebNLG

For the WebNLG sample, Dušek and Kasner (2020) created reference SED labels by mapping human quality judgements on a 1–3 scale (crowdsourced for WebNLG 2017) to *OK* (>=2.5) and *not OK* (<2.5). The crowdscourced quality judgements exist for a subset of 223 inputs from the WebNLG 2017 test set each paired with 10 different NLG outputs from participating systems. SED labels were taken to differ unless they were both *OK*, or one was *not OK* and the other was one of *omission, hallucination, omission+hallucination*).[7]

In the sample Dušek and Kasner annotated in their error analysis, the counts for reference labels derived from human ratings and for the NLI-SED system generated labels look as shown in the lower half of Table 1. The *not OK* label count of 46 shown for the NLI-SED system breaks down into 29 cases of omission, 13 cases of hallucination, and 4 cases of combined omission+hallucination).

In Dušek and Kasner's annotations, the reference label (the mapped human rating) was deemed to be correct (and the NLI-SED system prediction wrong) in 51 out of 100 cases, and the NLI-SED system prediction was deemed correct (and the reference label wrong) in 42 cases. In 7 cases either both reference label and NLI-SED system prediction were deemed wrong or the evaluators were unable to decide. These numbers are also shown in the top part of the second Dusek & Kasner column in Table 2)

The five error class labels for the WebNLG error annotations were as shown below. Each item may have more than one or none of these. The first three classes indicate, where possible, the likely source of the error in the SED label that was deemed wrong (produced by either the NLI-SED system or the mapped human ratings). Otherwise one of the last two will apply. Label frequencies are shown in lower part of the second Dusek & Kasner column in Table 2. NB: each error class predominantly applies to the underlined method.

1. **NLI-SED system error due to poor triple-to-text input mapping ('biased template') [bias-templ]**: Incorrect NLI-SED system label due to an inappropriate template being used in mapping the input triples to text (templates tend to work better for certain subject/object values, but the same template is used for all cases with a given predicate), resulting in ungrammatical sentences or even shift in meaning.
2. **NLI-SED system failure to recognise subject or object semantic equivalence ('value format') [val-format]**: In the verbalisation, the formatting of a subject or object differs from the input to the extent where the NLI check in the NLI-SED system failed to recognise them as equivalent in meaning (e.g. metres vs. kilometres).
3. **Incorrect reference SED label due to disfluent verbalisation ('bad sentence') [bad-sent]**: The human reference label, mapped to *not OK*, is incorrect, and this is likely because the human rating was affected by the disfluency/ungrammaticality of the verbalisation.
4. **Other cases of incorrect *OK* label ('unjustified OK') [unj-OK]**: The incorrect label (from either the human references or the NLI-SED system) is OK, and none of the above apply.
5. **Other cases of incorrectly identifying a semantic error ('unjustified not OK') [unj-notOK]**: the incorrect label (from either the human references or the NLI-SED system) either literally a *not OK* label, or one of *omission, hallucination, omission+hallucination*, *not OK*, and none of the above apply.

2.3 Reproduction targets

In the present context, there are four types of results that are candidates for reproduction: (i) *single numeric values* for the same measure (e.g. the overall number of times the SED label produced by the NLI-SED system was correct); (ii) *sets of numeric values for a set of related measures* (e.g. the numbers of input/output pairs annotated with each error label); (iii) *sets of discrete labels* from the same task (e.g. the correct/incorrect labels assigned to the NLI-SED system labels and the reference SED labels); and (iv) *unstructured textual comments* from the same task (here, the evaluator notes for each of the SED-label error annotations).

In order to draw conclusions regarding repeatability and reproducibility, results from original and

[7] One case of agreement, where the mapped human label was *Not OK* and the NLI-SED system produced *omission*, was included by mistake.

reproduction studies need to be compared, and how they're compared depends on which type (*i*, *ii*, *iii*, or *iv* above) a result is. We pick this up again in Section 3; here we list the results of types (i)–(iii) from Dušek and Kasner that we attempt to reproduce in our two reproduction studies (the free textual comments (type *iv*) were too disparate for us to try to compare):

i. Single numeric values (overall counts):
   a. Count of reference correct;
   b. Count of NLI-SED system correct;
   c. Count of both reference and NLI-SED system incorrect or evaluators couldn't decide;
   d. Count of individual error labels, six different labels for E2E, five for WebNLG (see Tables 2 and 3 for short-form labels).

ii. Sets of related numeric values:
   a. Set of counts of *Correctness* labels (i.a–i.c above);
   b. Set of counts of SED *Error-class* labels (i.d above).

iii. Sets of categorical values:
   a. Set of *Correctness* labels (one of {*NLI-SED, reference, neither*}; exactly one label per evaluation item);
   b. Set of SED *Error-class* labels; multiple labels per evaluation item).

3 Approach to Reproduction

For results of type *i* above (where we have single measured quantity values to compare), we follow the quantified reproducibility assessment (QRA) approach (Belz et al., 2022) which means (a) identifying and documenting (as we do in the attached HEDS sheet) the properties of evaluation experiments as standardised attribute-value pairs (*conditions of measurement* in QRA terms); and (b) computing the small-sample coefficient of variation (CV^*) over compared quantity values, as the measure of degree of reproducibility. QRA assessment results are shown in Tables 2 and 3 and discussed in Section 4.1.

For type *ii* results (Table 4, Section 4.2) we compute Pearson's r for pairwise correlation.

For results of type *iii*, we compute Fleiss's kappa (the multi-evaluator generalisation of Scott's pi) on aligned sets of categorical values where we have exactly one label per item (which is the case for the correctness labels), and Krippendorff's alpha where we have multiple labels per item (which is the case for the error labels). Results are shown in Table 5 and discussed in Section 4.3.

4 Two Reproductions

Our two reproduction studies repeated the Dušek and Kasner evaluations as closely as possible, the first using the same evaluators, the second using different evaluators. There were two complicating factors, necessitating the use of (i) new evaluator instructions, and (ii) a different way of allocating evaluators to evaluation items.

The reason for the difference in evaluator instructions is that in the original work, instructions were not written down, a shared understanding being evolved in the course of the work instead. In order to repeat the evaluations with new evaluators less familiar with the work, instructions had to be written down and shared which were then used in all reproductions. The instructions are included verbatim in the appendix.

Regarding evaluator allocation, in the original work, the work was shared between the two authors who each did about half of E2E and half of WebNLG, but it was not recorded who did which ones. For that reason, we decided to get the evaluators in the reproduction studies (the original two authors, and authors 4 and 5 of this paper) to each annotate all 100 E2E items and all 100 WebNLG items, and then we randomly selected half from each evaluator pair for a like-for-like comparison (in the tables below we call this the *combined* set of annotations). Assessing the similarity between these combined results and the original results forms the main body of our reproduction study: type *i* results are shown in Table 2, type *ii* in Table 4, and type *iii* in Table 5.

Additionally, we compare the four complete sets of annotations with the original annotations individually, for the single numeric values (type *i* results) from E2E and WebNLG only (Table 3).

Each evaluator worked on a separate Google spreadsheet in the exact same format as in the original study,[8] except that in the repeatability test which involved the original annotators, we shuffled the order of evaluation items to avoid inadvertent

[8] A blank copy of the evaluation sheet can be found here: https://docs.google.com/spreadsheets/d/1_4DZVu6Ow-9kZOjQJCjg2qZCLUt435Og

recall of original annotations.

4.1 Comparison of type *i* results

The results from the QRA test on label counts (type *i* results, i.e. single numeric values) for the *combined* annotations are shown in Table 2. The counts from the original study are in the Dusek & Kasner column in the left half of the table for E2E,[9] and in the right half for WebNLG. Counts from the repeatability (original annotators) and reproducibility (new annotators) tests and the corresponding CV* scores are shown in columns labeled as such in each half.

Looking at correctness label counts for E2E (rows 1–3, left half), the original annotators (A1+A2) are on the whole better able to reproduce their own results than the new annotators (A3+A4), which is as expected. However, if we look at the corresponding figures for WebNLG (right half) it turns out that here, the *new* annotators reproduce the original counts more closely. In terms of differences between correctness labels, the 'SED correct' counts are overall easiest to reproduce.

Moving on to error class counts, for E2E, CV* is broadly the same for original/new annotators for error classes relating to specific slots (eatType, priceRange, famFriend), but considerably worse for the new annotators for the remaining, more generic, error classes. For WebNLG, it is a more mixed picture: the new annotators reproduce the original counts better than the original annotators for error classes val-format and bad-sent, worse for error classes bias-templ and unj-OK, and equally well for error class unj-notOK.

Table 3 sheds additional light on the reproducibility of the individual category counts, by looking at the larger sets of 400 new annotations compared to the original 100, for each of E2E and WebNLG, thus providing a larger sample for, and more reliable estimates from, CV*. The two halves of the table are structured as in Table 2.

The results in Table 3 provide overall estimates across all five sets of annotations of the degree of reproducibility of the individual types of counts. For both E2E and WebNLG, correctness label counts are far easier to reproduce than error class counts which is as expected. Beyond that, again the 'SED correct' count is the most reproducible for both E2E and WebNLG. For E2E, counts for errors related to the priceRange slot (priceRange) are easiest to reproduce, whereas false negative hallucinations (f-halluc) are by far the hardest. For WebNLG, counts for bad-sent (bad grammar/fluency likely leading to 'not OK' label) are easiest to reproduce, and counts for val-format (phrases that are semantically equivalent not being recognised as such) are by far the hardest.

To put these CV* numbers into perspective, in the first ReproGen Shared Task, all except one (an outlier above 70) of the CV* scores for human evaluations were below 39 (Belz et al., 2020).

4.2 Comparison of type *ii* results

The results in the preceding section showed how reproducible correctness and error label *counts* were, for each count type independently, and regardless of whether labels were attached to the same items. In this section, we look at sets of counts in conjunction, and in the next section we look at labels as attached to evaluation items. Table 4 presents results from correlation tests on the set of all three correctness labels (top half), and on the set of all five (WebNLG) or six (E2E) error labels (sets of related numeric values, i.e. type *ii* results). Here too we are using the *combined* annotations.

We can see from the Pearson's r values that for both E2E and WebNLG all correlations are strong for the sets of correctness label counts. For the error class label count sets, on the other hand, while the original annotators achieve high correlation with their own earlier label counts for E2E, they do not for WebNLG, where the correlation is weak. The correlation between the A1+A2 and A3+A4 error label counts is weak to medium for both E2E and WebNLG. The new annotators do a reasonable job reproducing the original labels for E2E (r=0.62), but worst by far is the pronounced negative correlation for the new annotators for the WebNLG error labels.

4.3 Comparison of type *iii* results

The results from the agreement tests with Fleiss's kappa and Krippendorff's alpha on both label types as attached to evaluation items (type *iii* results, i.e. related sets of categorical values), again on the *combined* annotations, are shown in Table 5. For E2E and correctness labels, a similar picture emerges as previously in that agreement is similarly good across all comparisons, reflected also in the '%=' column which shows the percentage of times there was perfect agreement across all labels and all an-

[9] Counts for *ref correct*, *SED correct*, and *other* do not add up to 100 because of 3 missing annotations.

Counts of	E2E						Counts of	WebNLG					
	D&K	A1	A2	A3	A4	*CV**		D&K	A1	A2	A3	A4	*CV**
ref correct	34	41	31	37	50	21.325	ref correct	51	43	34	55	48	19.598
SED correct	45	45	53	41	47	10.594	SED correct	42	44	30	37	48	19.291
other	18	14	15	22	3	55.016	other	7	12	13	8	4	46.984
[eatType]	5	10	5	2	8	57.382	[bias-templ]	22	18	16	7	2	70.856
[priceRange]	30	31	39	42	9	47.756	[val-format]	7	1	3	26	0	162.088
[famFriend]	10	11	10	8	1	56.718	[bad-sent]	14	27	15	9	6	63.275
[f-halluc]	8	8	3	38	0	149.505	[unj-OK]	8	31	17	48	0	102.418
[f+omiss]	16	10	14	42	6	89.937	[unj-notOK]	15	16	25	26	1	67.727
[f+halluc]	17	15	24	19	4	52.288							

Table 3: QRA assessment of individual numeric results (type *i*), using the 4 sets of ***individual*** annotations.

	Pearson's r	E2E	Web-NLG
Correctness	Orig vs. A1+A2	0.999	0.965
	Orig vs. A3+A4	0.948	0.963
	A1+A2 vs. A3+A4	0.959	0.857
Error classes	Orig vs. A1+A2	0.947	0.209
	Orig vs. A3+A4	0.620	-0.630
	A1+A2 vs. A3+A4	0.373	0.414

Table 4: Pearson's r for counts of correctness and error-class labels (type *ii*), using the ***combined*** annotations (see Table 2 caption and Section 4).

			E2E	% =	Web-NLG	% =
Correctness	Fleiss's κ	All	0.674	71%	0.269	40%
		Orig vs. A1+A2	0.676	81%	0.140	50%
		Orig vs. A3+A4	0.677	81%	0.527	73%
		A1+A2 vs. A3+A4	0.643	78%	0.112	48%
Error classes	Kripp.'s α	All	0.467	12%	0.165	3%
		Orig vs. A1+A2	0.735	60%	0.207	21%
		Orig vs. A3+A4	0.347	15%	0.114	7%
		A1+A2 vs. A3+A4	0.330	18%	0.166	12%

Table 5: Fleiss's kappa for correctness and Krippendorff's alpha for error-class labels (type *iii*), using the ***combined*** annotations (see Table 2 caption and in text). '% =' = percentage of items with identical labels.

notators in a given comparison. For E2E and error class labels, the original annotators have strong agreement with their own original annotations, and the rest of the comparisons show medium agreement.

Again the picture is more mixed for WebNLG, where the new annotators have medium label-level agreement with the original labels for correctness, but for the other seven comparisons, label-level agreement is quite startlingly low (0 being chance).

5 Discussion

The error-analysis based evaluation method in this paper compares system outputs with reference outputs, but rather than just counting it against the system if there is disagreement between the two, it examines which is actually right in each case, also identifying the types of errors made by each. For E2E, 4 out of 5 sets of annotations (Table 3) agreed that the NLI-SED system was more often correct than the (automatically generated) references; for WebNLG the balance was slightly tipped in favour of the references (here derived from human ratings). These were important findings in the original paper, and are confirmed in all reproductions.

Other broad-strokes findings that are confirmed in all reproductions are that errors to do with priceRange predicate are the most common, and errors connected to eatType and famFriend are the least common, of the errors considered in E2E. For the error labels in WebNLG no findings are supported by all sets of annotations.

The degree to which the different types of results were reproducible varied. The more high-level correctness labels saw far better agreement than the more fine-grained error labels which also involve greater cognitive load. Moreover, the different backgrounds of the annotators and their degree of familiarity with the system and data may also have contributed to variation.

It is likely that if our instructions had been more precise, and more training/discussions of annotators in interpreting the instructions had taken place, the variation between studies would have been lower, and we can see room for improvement in this respect which we plan to explore in future work, where we will aim to:

- Ensure that annotators are given all relevant information for fully informed assessment of all error categories.
- Follow the iterative cycle in designing a linguistic annotation scheme (Pustejovsky et al., 2017): start with a preliminary annotation scheme and iteratively improve it using empirical observations (Howcroft et al., 2020).
- After a good fit between annotation scheme and task has been achieved and annotators reach a shared understanding, explicitly write down the annotation guidelines including any conclusions from informal discussions.

The iterative annotation design and written guidelines would have been useful even for repeating the study with the original annotators, as even their annotations differed in the repeat. We also noted some ideas for improving the error classes, which probably would have been already implemented with an iterative approach.

6 Conclusion

In this paper, we described two reproductions of a manual error analysis of the outputs from a semantic error detection (SED) system based on two-way entailment detection by an NLI model. We selected three types of results for reproduction, namely single numeric values, sets of numeric values, and sets of discrete labels, each of which requires different methods of comparison. All three types of results have broadly similar degrees of reproducibility: higher-level findings are mostly confirmed but lower-level agreement measures show a more differentiated picture, and are particularly low for WebNLG and error classes. Results for E2E are generally better reproduced than WebNLG, and correctness labels are easier to reproduce than the more fine-grained error classes.

In terms of conclusions to be drawn from the reproduction studies reported here, as with many other reproductions we found that the details of design and execution of the original study had not been recorded at the level of detail required for a reproduction. As a field, NLP is not currently in the habit of recording design/execution details of human evaluations very comprehensively or testing reproducibility during method development, for which time and other resources are often cited as reasons. The latter would be mitigated by the use of standard methods and tools for recording details of experiments and for assessing reproducibility.

Acknowledgements

Huidrom's work on this project was supported by the Faculty of Engineering and Computing, DCU, via a PhD grant. Belz's work was in part funded by EPSRC Grant No. EP/V05645X/1 for the ReproHum project. Dušek and Kasner's work was supported by ERC Grant No. 101039303 NG-NLG. Dušek's work is additionally supported by the Czech Ministry of Education project No. LM2018101 LINDAT/CLARIAH-CZ. Kasner's work is additionally supported by Charles University projects GAUK 140320 and SVV 260575.

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

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 (ACL'22)*.

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.

António Branco, Nicoletta Calzolari, Piek Vossen, Gertjan Van Noord, Dieter van Uytvanck, João Silva, Luís Gomes, André Moreira, and Willem Elbers. 2020. A shared task of a new, collaborative type to foster reproducibility: A first exercise in the area of language science and technology with REPROLANG2020. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5539–5545, Marseille, France. European Language Resources Association.

K. Bretonnel Cohen, Jingbo Xia, Pierre Zweigenbaum, Tiffany Callahan, Orin Hargraves, Foster Goss, Nancy Ide, Aurélie Névéol, Cyril Grouin, and Lawrence E. Hunter. 2018. Three dimensions of reproducibility in natural language processing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with

natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 131–137.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge. *Comput. Speech Lang.*, 59(C):123–156.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. 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.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

James Pustejovsky, Harry Bunt, and Annie Zaenen. 2017. Designing annotation schemes: From theory to model. In *Handbook of Linguistic Annotation*, pages 21–72. Springer.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Appendix: Annotator Instructions

A.1 Terms and Abbreviations

- **Semantic Error Detection (SED)**: in data-to-text generation, the task of deciding which errors if any are present in the output relative to the input.
- **SED label**: a label produced by an SED method indicating the semantic error class; typical label set e.g. Ok, omission, hallucination, omission+hallucination.
- **SED method**: here, one of D&K NLI-SED system, slot-error script, the human reference labels from WebNLG.
- **E2E slot**: an attribute in an E2E input, e.g. eatType=?.
- **Template**: a short template for converting each triple to text, used for the NLI checks (links for the lists of templates can be found here: E2E, WebNLG).

A.2 Instructions

A.2.1 E2E and WebNLG

First examine the input/output pair and make a note in the 'Other' column indicating the likely source of the error made by the incorrect SED method(s). Then, choose one or more of the error classes below that match the note. If none match, leave empty.

A.2.2 E2E

Each class indicates the likely source of the error made by the SED method that was deemed wrong (here, either the NLI-SED system or the slot-error script), and as many of the labels should be selected as apply to each item, in some cases none. NB: each error class predominantly applies to the underlined method.

1. *Error related to eatType=restaurant slot value*: The incorrect SED label (produced either by the NLI-SED system or the slot error script) is likely caused by something involving the slot/value pair eatType=restaurant, e.g. not detecting a hallucination when the eatType slot is not in the input, but the output mentions a restaurant.
2. *Error related to priceRange slot*: The incorrect SED label (produced by either the NLI-SED system or the slot error script) makes

an error related to the priceRange slot, e.g. incorrectly identifying a hallucination in the priceRange slot, when the price range information has in fact been correctly verbalised.

3. *Error related to familyFriendly attribute*: The incorrect SED label (produced by either the NLI-SED system or the slot error script) makes an error related to the familyFriendly slot, e.g. incorrectly identifying an omission when the information has in fact been correctly verbalised.

4. *Other false negative hallucination ('off-topic blabber')*: The incorrect SED label (produced by either the NLI-SED system or the slot error script) fails to detect a hallucination (unrelated to E2E slots) present in the verbalisation.

5. *Other false positive omission ('unjustified omission')*: the incorrect SED label (produced by either the NLI-SED system or the slot error script) wrongly detects an omission in the verbalisation.

6. *Other false positive hallucination ('unjustified hallucination')*: the incorrect SED label (produced by either the NLI-SED system or the slot error script) wrongly detects a hallucination in the verbalisation.

A.2.3 WebNLG

Each item may have more than one or none of these. The first three classes indicate, where possible, the likely source of the error in the SED label that was deemed wrong (produced by either the NLI-SED system or the reference SED label mapped from the human scores). Otherwise one of the last two will apply. Label frequencies are shown in the second Dušek & Kasner column in Table 1).

1. *SED system error due to poor triple-to-text input mapping ('biased template')*: incorrect NLI-SED system label due to an inappropriate template being used in mapping the input triples to text (templates tend to work better for certain subject/object values, but the same template is used for all cases with a given predicate), resulting in ungrammatical sentences or even shift in meaning. NB: please refer to the WebNLG templates as necessary.

2. *NLI-SED system failure to recognise subject or object semantic equivalence ('value format')*: in the verbalisation the formatting of a subject or object differs from the input to the extent where the NLI check in the NLI-SED system failed to recognise them as equivalent in meaning (e.g. metres vs. kilometres).

3. *Incorrect reference SED label due to disfluent verbalisation ('bad sentence')*: the incorrect human reference is not OK, and this is likely because the human rating was affected by the disfluency/ungrammaticality of the verbalisation.

4. *Other cases of incorrect OK label ('unjustified OK')*: the incorrect label (from either the human references or the NLI-SED system) is OK, and none of the above apply.

5. *Other cases of incorrect not OK label ('unjustified not OK')*: the incorrect label (from either the human references or the NLI-SED system) is not OK, and none of the above apply.

B Appendix: HEDS-Light Datasheet

Link to our HEDS Datasheet.