ReproGen at INLG 2022

TWO REPRODUCTIONS OF A HUMAN-ASSESSED COMPARATIVE EVALUATION OF A SEMANTIC ERROR DETECTION SYSTEM

Rudali Huidrom, Adapt/DCU
Ondrej Dusek, Charles University
Zdenek Kasner, Charles University
Thiago Castro Ferreira, UFMG
Anya Belz, Adapt/DCU And University Of Aberdeen.

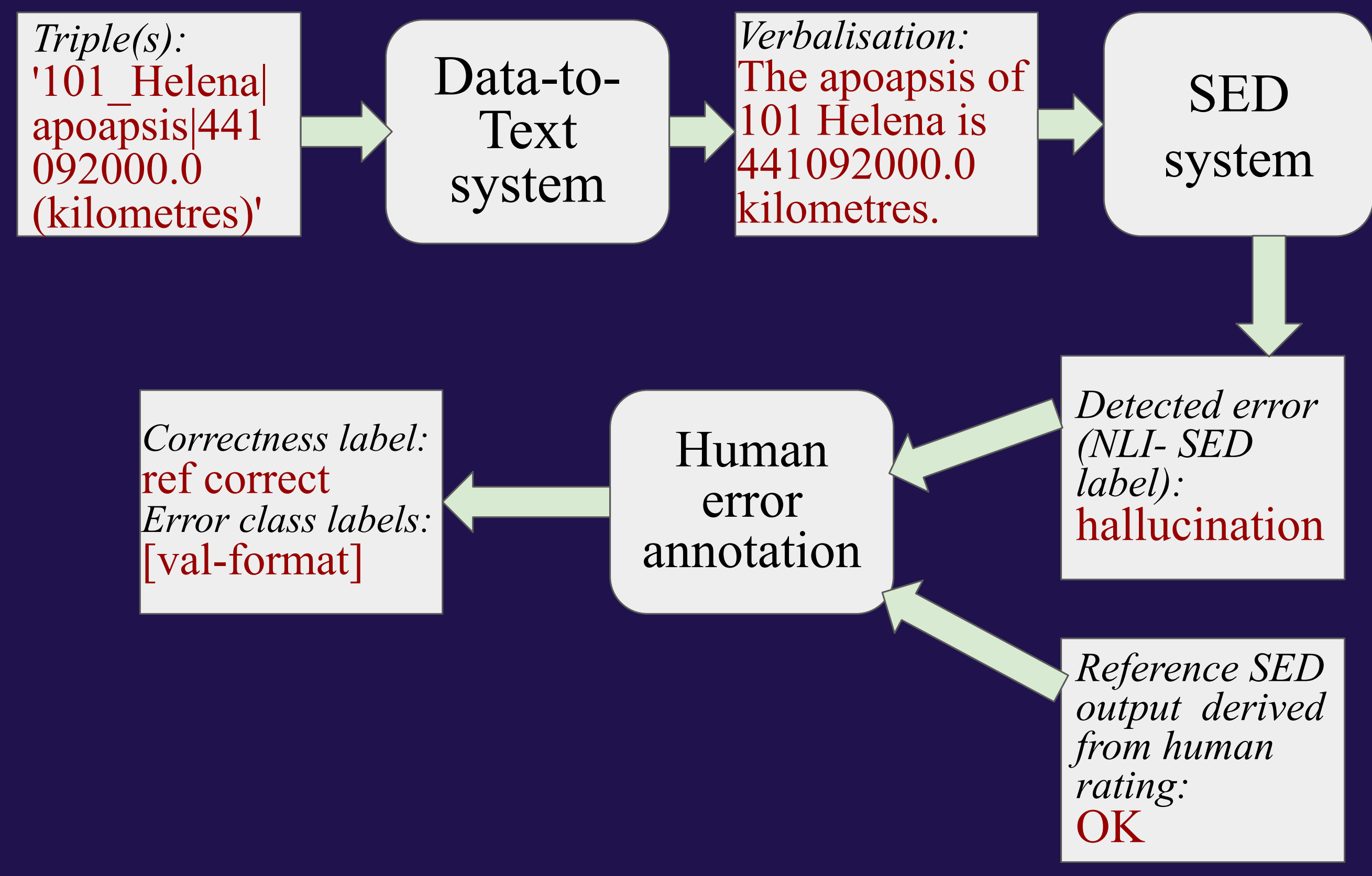
1. STUDY AIM

Two reproduction studies for the human evaluation originally reported by Dušek and Kasner (2020) in which the authors comparatively evaluated outputs produced by a *Semantic Error Detection (SED)* system for Data-To-Text Generation against reference outputs.

2. TWO REPRODUCTIONS

- In the first study, 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.

3. EXAMPLE


Triple(s): '101_Helena|apoapsis|441092000.0 (kilometres)'

Data-to-Text system

Verbalisation: The apoapsis of 101 Helena is 441092000.0 kilometres.

SED system

Detected error (NLI- SED label): hallucination

Reference SED output derived from human rating: OK

Human error annotation

Correctness label: ref correct
Error class labels: [val-format]

4. MANUAL EVALUATION OF THE SED METHOD

E2E & WebNLG (correctness labels)

(a) Counts of reference labels [ref correct].
(b) Counts of NLI-SED generated system labels [SED correct].
(c) Either (a) and (b) are wrong or the evaluators can't decide [other].

E2E (error class labels):

- Error related to eatType=restaurant slot value [eatType]
- Error related to priceRange slot [priceRange]
- Error related to familyFriendly attribute [famFriend]
- Other false negative hallucination ('off topic blabber') [f-halluc]
- Other false positive omission ('unjustified omission') [f+omiss]
- Other false positive hallucination ('unjustified hallucination') [f+halluc]

WebNLG (error class labels):

- Poor triple-to-text input mapping ('biased template') [bias-templ]
- Failure to recognise subject or object semantic equivalence ('value format') [val-format]
- Incorrect SED label due to disfluent verbalisation ('bad sentence') [bad-sent]
- Other cases of incorrect OK label ('unjustified OK') [unj-OK]
- Other cases of identifying a semantic error ('unjustified not OK') [unj-notOK]

5. REPRODUCTION TARGETS

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 the evaluators couldn't decide.
  d. Count of individual error labels, six for E2E and five for WebNLG.

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 team)
  b. Set of SED Error class labels (multiple labels per evaluation team).

6. APPROACH TO REPRODUCTION

- For type i results, we follow Quantified Reproducibility Assessment, QRA (Belz et al., 2022).
- For type ii results, we compute Pearson's r for pairwise correlation.
- For type iii results, we compute Fleiss' kappa on aligned categorical values where we have exactly one label per item (correctness labels) and Krippendorff's alpha where we have multiple labels per item (error class labels).

7. QRA OF CORRECTNESS / ERROR LABEL COUNTS FOR NON-COMBINED & COMBINED ANNOTATIONS (TYPE I RESULTS)

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							

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						

8. PEARSON'S R FOR TYPE II RESULTS; FLEISS'S KAPPA ON CORRECTNESS LABELS, KRIPPENDORFF'S ALPHA FOR ERROR-CLASS LABELS FOR TYPE III RESULTS, FOR REPEATABILITY TEST (A1+A2) AND REPRODUCIBILITY TEST (A3+A4).

	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

			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%

9. FINDINGS

- Type i results: original annotators reproduce *correctness label* counts more closely than new annotators for E2E. For WebNLG, new annotators reproduce *correctness label* counts more closely. Reproduction of *error class label* counts is broadly the same for both sets of annotators for E2E, whereas for WebNLG, it is a mix.
- Type ii results: correlation is high for *correctness labels for* both E2E and WebNLG. For *error class labels*, correlation is higher for original annotators in the case of E2E but not in the case of WebNLG.
- Type iii results: for E2E and *error class labels*, the annotators have strong agreements whereas it is more mixed for WebNLG.

10. IMPROVING REPRODUCIBILITY

- 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).
- Explicitly write down the annotation guidelines including any conclusions from informal discussions after a good fit between annotation scheme and task has been achieved and annotators reach a shared understanding

✉ rudali.huidrom@adaptcentre.ie
https://github.com/RHuidrom/reprogen22_dusek_and_kasner_2020

WE WELCOME ALL FEEDBACK AND PLEASE FEEL FREE TO GET IN TOUCH WITH US.