Supplementary Material: Evaluating Semantic Accuracy of Data-to-Text Generation with Natural Language Inference

Dataset details

The WebNLG data used for our experiments is the subset of NLG system outputs used for human evaluation (Shimorina et al., 2019)[7] – 223 sampled data inputs from the WebNLG 2017 test set with 10 different NLG system outputs for each input, i.e., 2,230 instances in total.

For the main E2E experiments (with the slot error script as ground truth in Section 5.2), we used the full set of primary system outputs on the whole E2E test set (Dušek et al., 2020) – 21 outputs for 630 input data items each, i.e., 13,230 instances in total.[8] For semantic classifier comparison in Section 5.3, we used two expert-annotated sets provided to us by Dušek et al. (2019), who used them for their slot error script evaluation: 200 instances from the E2E training set (the human-written texts were reannotated as many of them did not reflect the original input properly) and 400 outputs of different variants of the TGen NLG system (Dušek and Jurčíček, 2016).

WebNLG Error Analysis

We checked 100 randomly sampled examples from the WebNLG data where our approach and the crowdsourced human annotation gave different results. We identified 51 cases where the crowdsourced human annotation was indeed correct, 42 where the human annotation was incorrect but our NLI-based approach provided a correct result, and 7 cases where both annotations were incorrect or where it was not possible to unambiguously decide between the two.

The main error types identified, with counts and examples, are as follows (note that the analyzed examples may belong to multiple or none of the above classes):

(1) Annotation problems – 22 counts. Example:

Data
1 Decembrie 1918 University | state | Alba

Templates
1 Decembrie 1918 University stands in the state of Alba.

Text
1 decembrie 1918 university is in the state of alba.

Human Output
2.33 *(=not_OK)*

NLI Output
OK

Commentary
The sentence is OK, but the human score is slightly below the threshold for no apparent reason.

(2) Inaccurate templates – 22 counts. Example:

Data
Aenir | language | English language

Templates
One of the languages of Aenir is English language.

Text
aenir is written in english.

Human Output
3 *(=OK)*

NLI Output
hallucination

Commentary
The sentence is OK, but the template is not specific enough for a literary work, which leads the NLI to assume this is a hallucination.

[7] The data is available at https://gitlab.com/webnlg/webnlg-human-evaluation. We used the file all_data_final_averaged.csv.

[8] The data are available from https://github.com/tuetschek/e2e-eval. We used all the files under system_outputs/primary.

(3) Disfluent or disorganized sentences – 14 counts. Example:

Data	**Templates**
Spain | leaderName | Felipe VI of Spain	the name of the leader of Spain is Felipe VI of Spain.
Ajoblanco | region | Andalusia	Ajoblanco can be found Andalusia.
Andalusia | leaderName | Susana Díaz	the name of the leader of Andalusia is Susana Díaz.
Ajoblanco | country | Spain	Ajoblanco is situated within Spain.
Spain | demonym | Spaniards	The demonym for natives of Spain is Spaniards.

Text

ajoblanco is a spanish dish that originates andalusia , where the leader is susana diaz. felipe vi is the leader of spain , where the people that live there are called spaniards.

Human Output	**NLI Output**
2.33 *(=not_OK)*	*OK*

Commentary

The sentence is OK, but not completely fluent and not very well organized (some of the facts are not very related). This probably leads to the lower human score, but doesn't cause any problems to the NLI.

(4) Value format errors – 7 counts. Example:

Data	**Templates**
Asterix (comicsCharacter) | creator | René Goscinny	The creator of Asterix (comicsCharacter) is René Goscinny.
René Goscinny | nationality | French people	René Goscinny was French people.
Asterix (comicsCharacter) | creator | Albert Uderzo	The creator of Asterix (comicsCharacter) is Albert Uderzo.

Text

asterix was created by rené goscinny and albert uderzo, the former being a french national.

Human Output	**NLI Output**
2.5 *(=OK)*	*omission*

Commentary

The sentence is OK, but the format of the values is not maintained in the text ("comicsCharacter" is missing). The NLI treats this as an omission. Similar cases involve also e.g. number formatting.

E2E Error Analysis

We checked 100 randomly sampled examples from the E2E data where our approach and the slot error script gave different results. We identified 34 cases where the slot error script was indeed correct, 45 where the script was incorrect but our NLI-based approach provided a correct result, and 18 cases where both annotations were incorrect or where it was not possible to unambiguously decide between the two.

The main error types identified, with counts and examples, are as follows (note that the analyzed examples may belong to multiple or none of the above classes):

(1) Value interpretation problems – 40 counts. Example:

Data	**Templates**
The Punter | eat type | restaurant	The Punter is a restaurant.
The Punter | food | Indian	The Punter serves Indian.
The Punter | price range | high	The Punter is in the high price range.
The Punter | rating | average	The Punter has average customer rating.
The Punter | area | city centre	The Punter is located in the city centre.
The Punter | family friendly | no	The Punter is not family-friendly.
The Punter | near | Express by Holiday Inn	The Punter is located near Express by Holiday Inn.

Text

The Punter is a high priced, average rated, adult only Indian restaurant located near Express by Holiday Inn in the city centre.

Slot Error Script	**NLI Output**
OK	*hallucination*

Commentary

The text uses “adult only” to verbalize *family-friendly=no*, which is generally considered correct in the E2E dataset. However, the NLI treats this as hallucination (“adult only” does not necessarily follow from “is not family friendly”).

(2) Incorrect patterns in the slot error script – 33 counts. Example:

Data	**Templates**
The Cricketers | eat type | restaurant	The Cricketers is a restaurant.
The Cricketers | food | Chinese	The Cricketers serves Chinese.
The Cricketers | price range | cheap	The Cricketers is in the cheap price range.
The Cricketers | rating | average	The Cricketers has average customer rating.
The Cricketers | area | riverside	The Cricketers is located in the riverside.
The Cricketers | family friendly | yes	The Cricketers is family-friendly.
The Cricketers | near | All Bar One	The Cricketers is located near All Bar One.

Text

Cheap Chinese food for all the family can be found at The Cricketers restaurant, near All Bar One, in the riverside area. Average ratings.

Slot Error Script	**NLI Output**
omission	*OK*

Commentary

The slot error script considers *family-friendly=yes* as missing – it probably does not include the pattern “for all the family”. NLI has no problems handling this.

(3) Edge cases, hard to interpret – 18 counts. Example:

<table>
<tr><th>Data</th><th>Templates</th></tr>
<tr><td><i>The Mill | eat type | restaurant</i></td><td>The Mill is a restaurant.</td></tr>
<tr><td><i>The Mill | food | English</i></td><td>The Mill serves English.</td></tr>
<tr><td><i>The Mill | price range | moderate</i></td><td>The Mill is in the moderate price range.</td></tr>
<tr><td><i>The Mill | rating | 3 out of 5</i></td><td>The Mill has 3 out of 5 customer rating.</td></tr>
<tr><td><i>The Mill | area | riverside</i></td><td>The Mill is located in the riverside.</td></tr>
<tr><td><i>The Mill | family friendly | yes</i></td><td>The Mill is family-friendly.</td></tr>
<tr><td><i>The Mill | near | Café Rouge</i></td><td>The Mill is located near Café Rouge.</td></tr>
</table>

Text

The Mill is a moderate restaurant that serves English food. Yes it is kids-friendly. Its customer rating is 3 out of 5. It is located in the riverside area near Café Rouge.

Slot Error Script	NLI Output
omission	*OK*

Commentary

It is hard to interpret “moderate restaurant” as a correct verbalization of *price_range=moderate*. The NLI makes this assumption while the slot error script does not.

(4) Off-topic hallucinations – 8 counts. Example:

<table>
<tr><th>Data</th><th>Templates</th></tr>
<tr><td><i>Giraffe | eat type | restaurant</i></td><td>Giraffe is a restaurant.</td></tr>
<tr><td><i>Giraffe | food | English</i></td><td>Giraffe serves English.</td></tr>
<tr><td><i>Giraffe | area | riverside</i></td><td>Giraffe is located in the riverside.</td></tr>
<tr><td><i>Giraffe | family friendly | yes</i></td><td>Giraffe is family-friendly.</td></tr>
<tr><td><i>Giraffe | near | Rainbow Vegetarian Café</i></td><td>Giraffe is located near Rainbow Vegetarian Café.</td></tr>
</table>

Text

Giraffe is a beautiful restaurant close to the Rainbow Vegetarian Café. It is reasonably liked place serves English food and is children friendly.

Slot Error Script	NLI Output
omission	*hallucination+omission*

Commentary

While both the slot error script and the NLI detect the missing verbalization of *area*, NLI probably interprets “beautiful” and “reasonably liked” as hallucinations, while the slot error script is not able to detect such cases.