E2E Refined Dataset

Keisuke Toyama, Katsuhito Sudoh, and Satoshi Nakamura
Nara Institute of Science and Technology
Ikoma, Japan
{toyama.keisuke.tb5, sudoh, s-nakamura}@is.naist.jp

***Abstract*—As a well-known meaning representation (MR)-to-text dataset, the E2E dataset has been used by many studies in natural language generation. However, the dataset suffers from many deletion, insertion, and substitution errors in its MR-text pairs that affect the quality of MR-to-text system trained using the dataset. In this paper, we develop a refined dataset by fixing text and MR errors, applying text normalization, and giving extra annotations on the MR part. We release Python codes to convert the original E2E dataset to the refined one on GitHub.**

***Index Terms*—data-to-text, meaning representation, mr-to-text, natural language generation**

I. Introduction

Natural language generation (NLG) [1] is a generative process that produces natural written or spoken language from input data not limited to text. For example, machine translation or question answering is an NLG task that generates text from an unstructured textual input. Data-to-text is another NLG task that generates text from structured inputs such as concepts, tables, knowledge graphs, and resource description frameworks (RDFs). Meaning representation (MR)-to-text is one of the data-to-text tasks where MR is composed of a collection of pairs of a brief text passage and a corresponding MR with several attribute-value pairs, as shown in Table I. There are several well-known corpora of MR-to-text, Weather (generating weather reports from meteorological data) [2], RotoWire (generating summaries of sports matches from game statistics) [3], WikiBio (generating biographies from Wikipedia infobox) [4] and so on. The E2E dataset [5] in a restaurant recommendation domain, used in the E2E NLG Challenge [6], is one of the most popular datasets for MR-to-text. However, this dataset was developed by crowdsourcing and suffers from errors in MR-text pairs that affect the performance of MR-to-text models. In this paper, we aim to refine the E2E dataset by resolving errors and giving extra annotations. We fix errors in MR-text correspondences and remove irrelevant data samples from the dataset. We also provide additional annotations: *Number of sentences*, *MR order*, and *Sentence indexes* to control the generated text more precisely. We demonstrate that the refined dataset, called *E2E Refined Dataset*[1], improves MR-to-text performance.

II. E2E Dataset

The E2E dataset is made up of pairs consisting of sentences of restaurant recommendations in British English and an MR, which corresponds to the sentences, with eight attributes as shown in Table I. However, some MR-text pairs contain deletion, insertion, and substitution errors. For example, in the text part of Table II, the value “English” for the `food` attribute and the value “high” for the `customer rating` attribute are missing, the value “city centre” for the `area` attribute is wrongly added, and the value “pub” for the `eatType` attribute and the value “more than £30” for the `priceRange` attribute are wrongly replaced with “restaurant” and “low”, respectively.

The dataset should not contain such wrong data for effective control over the sentence content in MR-to-text. Despite the existence of updated versions of the E2E dataset that address errors, including the cleaned [7] and enriched versions [8], these revised datasets still contain deletion, insertion, and substitution errors. We found a certain number of errors, shown in Table III. To address this issue, we fixed the inaccuracies in the correspondence between MR-text and discarded unsuitable data samples. Moreover, we refined the E2E dataset by manually annotating the MR values to provide further constraints from the text part. As a result, the E2E refined dataset is

TABLE I
Example of the E2E dataset

MR	`name`	The Olive Grove
	`eatType`	pub
	`food`	(empty)
	`priceRange`	moderate
	`customer rating`	(empty)
	`area`	riverside
	`familyFriendly`	yes
	`near`	(empty)
Text	Moderately priced The Olive Grove pub is located on the riverside. It welcomes kids.	

TABLE II
Example of MR errors in the E2E dataset: **Bold** indicates deletion error, <u>Underline</u> indicates insertion error, and *Italic* indicates substitution error.

MR	`name`	Wildwood
	`eatType`	*pub*
	`food`	**English**
	`priceRange`	*more than £30*
	`customer rating`	**high**
	`area`	(empty)
	`familyFriendly`	(empty)
	`near`	(empty)
Text	Wildwood is a *restaurant* providing take-away deliveries in the *low* price range. It is located in the <u>city centre</u>.	

[1]The dataset and Python programs are available at https://github.com/KSKTYM/E2E-refined-dataset/

TABLE III
NUMBER OF MR LABELLING ERRORS IN EACH DATASET

Error type	E2E dataset [5]			Cleaned dataset [7]			Enriched dataset [8]		
	Training	Validation	Test	Training	Validation	Test	Training	Validation	Test
Deletion	10,931	1,096	1,315	23	1	1	1,262	145	89
Insertion	10,028	263	16	4,475	471	745	25,570	2,724	3,082
Substitution	9,290	794	945	5,795	616	666	4,172	413	395

obtained with 40,560, 4,489, and 4,555 samples in the train, validation, and test splits, respectively. Table VII shows an example from it.

III. TEXT REFINEMENT

The E2E dataset contains errors and discrepancies in the language used. We improved the quality of the text parts of the dataset by correcting errors and standardizing expressions as follows.

A. Error Correction

We refined various kinds of errors presented in the original E2E dataset. Table IV shows their examples. We focused on the following four error categories.

1) Indefinite Articles: We fixed any errors in the usage of the indefinite articles “a” and “an.”

2) Irregular MR Values: Every MR is designed to contain just one value per attribute or none. Nonetheless, there are instances where some MR data might contain two values for a single attribute. To maintain data integrity, we deleted any inapposite data from the dataset.

3) Overlaps: We deleted duplicated phrases within a sentence.

4) Typos: We identified and fixed more than 3,700 typographical errors.

B. Normalization

We normalized the text in the following six types of aspects.

1) British English: Given that the E2E dataset adheres to British English, we substituted words such as “neighbor”, “favor”, and “specialize” spelt in American style with their British counterparts: “neighbour”, “favour”, and “specialise”.

2) Capital Letters: We refined how to capitalize letters in MR values and text as follows:

- all values of the `name` and `near` attributes,
- the first letter of all values in the `food` attribute except “fast food,”
- the first letter of each sentence.

3) Currency Expressions: For ease of use, we standardized the currency unit as “£20” instead of using variations such as “20 quid”, “20lb”, “20gbp”, “20 pounds”, and so on.

4) Prices: The category of `priceRange` is defined as “lower than £20”, “£20-25”, “more than £30”, “cheap”, “moderate”, and “expensive”, as shown in Table VI. In this context, “£22” should be annotated as “£20-25.” However, to eliminate confusion, we used the label “£20-25” for all prices falling within that range, including values like “£22”, “£23”, “£24”, and “from £20 to £25.” This approach was also taken for the labels “lower than £20” and “more than £30.”

5) Quotation Marks: We replaced double quotation marks with single quotation marks.

6) Symbols: We standardized symbols such as commas, periods, and white spaces.

IV. MR REFINEMENT

The MRs in the original E2E dataset include labelling errors. We refined the MR labels as described in Section IV-A. We provided additional annotations for further controllable generation study regarding flexible content planning, as described in Section IV-B.

A. Labelling

Throughout the E2E dataset, we corrected MR labelling errors manually. Additionally, we replaced the value “high” with “expensive” for the `priceRange` attribute. Moreover, we added new labels for the `food` attribute, including “American”, “Canadian”, and “Thai.” Table VI lists all the refined labels.

B. Additional Annotations

1) MR Order: As shown in Table VII, we marked the order of the MR values mentioned in the corresponding text. In case of an empty MR value, we represented the order with a “0.”

2) Number of Sentences: In Table VII, we marked the total number of sentences in the text. The count of sentences was established by looking for periods (“.”) and question marks (“?”). As per the sample in the table, the section of the text contains two periods. Therefore, we set the count of sentences as “2.”

3) Sentence Indexes: We also provided annotations for the appearance of each MR value in the corresponding sentences as shown in Table VII. For instance, the phrases related to the value “riverside” for the `area` attribute and the value “yes” for the `familyFriendly` attribute are found in the first and second sentences, respectively. In cases where an MR value is empty, we set the index to “0.”

V. MR-TEXT PAIR REFINEMENT

To further enhance the dataset, we refined the MR-text pairs by removing repetitions and utilizing a strategy to convey certain values effectively in both the MR and text.

A. Deduplication

We excluded approximately 1,500 MR-text pairs from the dataset as a result of the deduplication process.

TABLE IV
EXAMPLES OF ERROR CORRECTION

Refinement type	Original text	Refined text
Indefinite article	Cocom is **a** average family friendly restaurant.	COCOM is **an** average family friendly restaurant.
Irregular MR values	**Clare Hall** is known for Fast food and coffee shop style bakeries although, customers only rate them average the **Clowns** are quite amusing.	(removed)
Overlap	The Golden Curry served English food, **is adult only**, is in the city centre, **is adult only**, has a customer rating of 5 out of 5 and is near the Café Rouge.	THE GOLDEN CURRY served English food, **is adult only**, is in the city centre, has a customer rating of 5 out of 5 and is near the CAFÉ ROUGE.
Typos	Moderately priced fast **found** can be found at Blue Spice in city centre.	Moderately priced fast **food** can be found at BLUE SPICE in city centre.

TABLE V
EXAMPLES OF NORMALIZATION

Refinement type	Original text	Refined text
British English	Cheap family **favorite**, The Twenty Two, near The Rice Boat in riverside, got a 5 out of 5.	Cheap family **favourite**, THE TWENTY TWO, near THE RICE BOAT in riverside, got a 5 out of 5.
Capital letters	Giraffe is kid friendly. **i**t is located near riverside	GIRAFFE is kid friendly. **I**t is located near riverside.
Currency expression	The Punter is a Japanese restaurant under **20 pounds**.	THE PUNTER is a Japanese restaurant under **£20**.
Prices	If you're looking for pub grub or Indian food, you could try The Plough. No you can't take your kids there but the prices are reasonable about **£24** for a meal. You'll find it near to Café Rouge.	If you're looking for pub grub or Indian food, you could try THE PLOUGH. No you can't take your kids there but the prices are reasonable about **£20-25** for a meal. You'll find it near to CAFÉ ROUGE.
Quotation marks	A highly rated coffee shop **"**The Punter**"** serving English food priced between £20 - £25 and is child friendly.	A highly rated coffee shop **'**THE PUNTER**'** serving English food priced between £20-25 and is child friendly.
Symbols	Wildwood, located near the city center**.,** is a low price pub.	WILDWOOD, located near the city centre**,** is a low price pub.

TABLE VI
ALL VARIATIONS OF MR VALUES IN THE E2E REFINED DATASET

Attribute	Number of variations	MR values (delexicalized)
`Name`	1	NAME
`eatType`	4	(empty), coffee shop, pub, restaurant
`food`	11	(empty), fast food, American, Canadian, Chinese, English, French, Indian, Italian, Japanese, Thai
`priceRange`	7	(empty), less than £20, £20-25, more than £30, cheap, expensive, moderate
`customer rating`	7	(empty), average, high, low, 1 out of 5, 3 out of 5, 5 out of 5
`area`	3	(empty), city centre, riverside
`familyFriendly`	3	(empty), no, yes
`near`	2	(empty), NEAR

B. Delexicalization

As the value for the `name` attribute and that for `near` attribute are always directly reflected in the sentences, we standardized the data by substituting these values in the MR values and sentences with special letters, "NAME" and "NEAR." We retained the original values for both attributes, which are required to generate proper sentences, even though the delexicalized data are beneficial to train MR-to-text models.

VI. EXPERIMENTS

We investigated the effect of the refinement by the following experiments.

A. Dataset

We used the original E2E dataset and the E2E refined dataset. For a fair comparison, we editted some values in the E2E refined dataset as follows:

- only the first letter of each word of `name` and `near` values are capitalized,
- the value "expensive" for the `priceRange` attribute is reverted to "high" (see Section IV-A).

We did not use additional annotations (described in Section IV-B) either.

B. Method

We used TGen[2] [9], an LSTM-based sequence-to-sequence model that utilizes an attention mechanism. It was the baseline method of the E2E NLG Challenge. We developed two models. The first model was trained using the original E2E dataset, while the second was trained using the E2E refined dataset.

C. Metrics

We used BLEU [10], NIST [11], METEOR [12], ROUGE_L [13], and CIDEr [14], which were used for the E2E NLG challenge[3], as evaluation metrics for both models.

[2]https://github.com/UFAL-DSG/tgen
[3]https://github.com/tuetschek/e2e-metrics

TABLE VII
EXAMPLE OF THE E2E REFINED DATASET (ORIGINAL SAMPLE IS SHOWN IN TABLE I).

	Attribute	Value	Order	Sentence index
MR	`name`	NAME (THE OLIVE GROVE)	2	1
	`eatType`	pub	3	1
	`food`	(empty)	0	0
	`priceRange`	moderate	1	1
	`customer rating`	(empty)	0	0
	`area`	riverside	4	1
	`familyFriendly`	yes	5	2
	`near`	(empty)	0	0
Number of sentences	2			
Text	Moderately priced THE OLIVE GROVE pub is located on the riverside. It welcomes kids.			
Text (delexicalized)	Moderately priced NAME pub is located on the riverside. It welcomes kids.			

TABLE VIII
RESULTS OF THE EXPERIMENTS

Dataset	BLEU(↑)	NIST(↑)	METEOR(↑)	ROUGE_L(↑)	CIDEr(↑)
E2E dataset	0.5462	7.6209	0.4103	0.6561	2.2448
E2E refined dataset	0.5581	7.8378	0.4252	0.6488	2.3865

The evaluation was conducted on the test set of the E2E refined dataset.

D. Results

As listed in Table VIII, the scores reveal that the model trained on the E2E refined dataset surpassed the performance of the model trained on the original dataset, except for ROUGE_L. These results suggest that the refined dataset contains more accurate label information, which ultimately led to improved performance.

VII. LIMITATIONS

Despite our efforts to adapt the E2E dataset for the progression of MR-to-text models, several constraints persist:

- As discussed in Section III-A, we deleted data with irregular MR values. Nonetheless, different formulations of MR-to-text problems might permit multiple values in more intricate scenarios.
- We currently disregard referring expressions, even though they are generally acceptable.
- We treat all attributes except `name` as modifiers of a `name`. However, there are instances where an attribute modifies `near`, which our current formulation overlooks.

VIII. CONCLUSION

In this study, we presented our E2E refined dataset. We reduced the number of errors in the original dataset by fixing inaccuracies and standardizing phrases. Furthermore, we added new annotations, number of sentences, MR order, and sentence indexes, enabling us to control the generated text more precisely. Our tests indicated that this refined dataset contributed to NLG's better performance. We expect this dataset to inform future research in various areas, including data-to-text.

ACKNOWLEDGMENT

Part of this work was supported by JSPS KAKENHI Grant Number JP21H05054.

REFERENCES

[1] C. Dong, Y. Li, H. Gong, M. Chen, J. Li, Y. Shen, and M. Yang, "A survey of natural language generation," ACM Computing Surveys, vol. 55, issue 8, pp. 1–38, 2022.

[2] A. Balakrishnan, J. Rao, K. Upasani, M. White, and R. Subba, "Constrained decoding for neural NLG from compositional representations in task-oriented dialogue," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 831–844, 2019.

[3] S. Wiseman, S. M. Shieber, and A. M. Rush, "Challenges in data-to-document generation," in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2253–2263, 2017.

[4] R. Lebret, D. Grangier, and M. Auli, "Neural text generation from structured data with application to the biography domain,", in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1203–1213, 2016.

[5] J. Novikova, O. Dušek, and V. Rieser, "The e2e dataset: new challenges for end-to-end generation," in Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, pp. 201–206, 2017.

[6] O. Dušek, J. Novikova, and V. Rieser, "Evaluating the state-of-the-art of end-to-end natural language generation: the e2e NLG challenge," Computer Speech and Language, vol. 59, pp.123–156, 2020.

[7] O. Dušek, D. M. Howcroft, and V. Rieser, "Semantic noise matters for neural natural language generation," in Proceedings of the 12th International Conference on Natural Language Generation, pp. 421–426, 2019.

[8] T. C. Ferreira, H. Vaz, B. Davis, and A. Pagano, "Enriching the e2e dataset," in Proceedings of the 14th International Conference on Natural Language Generation, pp. 177–183, 2021.

[9] O. Dušek and F. Jurčíček, "Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 45–51, 2016.

[10] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a method for automatic evaluation of machine translation," in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318, 2002.

[11] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in Proceedings of the 2nd International Conference on Human Language Technology Research, pp. 138–145, 2002.

[12] M. Denkowski and A. Lavie, "Meteor universal: language specific translation evaluation for any target language," in Proceedings of the 9th Workshop on Statistical Machine Translation, pp. 376–380, 2014.

[13] C. Lin, "ROUGE: a package for automatic evaluation of summaries," in Text Summarization Branches Out, pp. 74–81, 2004.

[14] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: consensus-based image description evaluation," in Proceedings of 2015 IEEE conference on Computer Vision and Pattern Recognition, pp. 4566–4575, 2015.