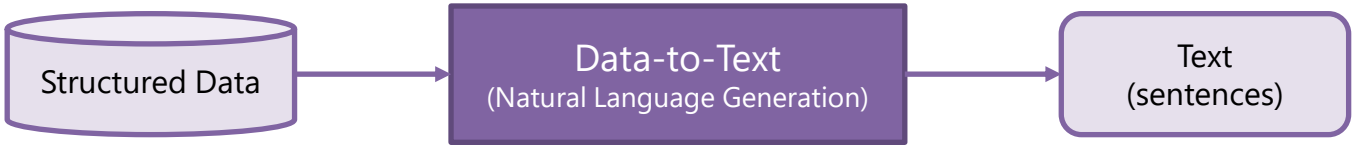
E2E Refined Dataset

Keisuke Toyama, Katsuhito Sudoh and Satoshi Nakamura
Nara Institute of Science and Technology

Data-to-Text



Weather

condition1	sunny
date_time1	this weekend
avg_high1	60s
low2	43
date_time2	Sunday evening
chance3	likely
wind_summary3	strong
date_time3	Saturday morning

It'll be sunny throughout this weekend. The high will be in the 60s, but expect temperatures to drop as low as 43 degrees by Sunday evening. There's also a chance of strong winds on Saturday morning.

WikiBio

Frederick Parker-Rhodes

Born	21 November 1914 Newington, Yorkshire
Died	2 March 1987 (aged 72)
Residence	UK
Nationality	British
Fields	Mycology, Plant Pathology, Mathematics, Linguistics, Computer Science
Known for	Contributions to computational linguistics, combinatorial physics, bit-string physics, plant pathology, and mycology
Author abbrev. (botany)	Park.-Rhodes

Frederick Parker-Rhodes (21 March 1914–21 November 1987) was an English linguist, plant pathologist, computer scientist, mathematician, mystic, and mycologist.

RotoWire

TEAM	WIN	LOSS	PTS	FG_PCT	RB	AS ...
Heat	11	12	103	49	47	27
Hawks	7	15	95	43	33	20

PLAYER	AS	RB	PT	FG	FGA	CITY ...
Tyler Johnson	5	2	27	8	16	Miami
Dwight Howard	4	17	23	9	11	Atlanta
Paul Millsap	2	9	21	8	12	Atlanta
Goran Dragic	4	2	21	8	17	Miami
Wayne Ellington	2	3	19	7	15	Miami
Dennis Schroder	7	4	17	8	15	Atlanta
Rodney McGruder	5	5	11	3	8	Miami
Thabo Sefolosha	5	5	10	5	11	Atlanta
Kyle Korver	5	3	9	3	9	Atlanta
...						

The Atlanta Hawks defeated the Miami Heat, 103-95, at Philips Arena on Wednesday. Atlanta was in desperate need of a win and they were able to take care of a shorthanded Miami team here. Defense was key for the Hawks, as they held the Heat to 42 percent shooting ...(*snip*)

E2E dataset

- **MR(Meaning Representation)-to-Text dataset**
  - MR: 8 attribute-value pairs
  - Text: British English sentences in the restaurant recommendation domain

	attribute	value
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.	

  - used in the E2E NLG Challenge (2017)

Deletion/Insertion/Substitution Errors in E2E dataset

- **E2E dataset includes many errors**

	attribute	value
MR	name	The Punter
	eatType	coffee shop
	food	***English***
	priceRange	***moderate***
	customer rating	1 out of 5
	area	***(empty)***
	familyFriendly	yes
	near	Cafe Sicilia
Text	The Punter is a ***cheap*** family friendly coffee shop located in ***City Centre*** near Cafe Sicilia. 1 out of 5 customer rating.	

English is missing
↓
deletion error

moderate is replaced with ***cheap***
↓
substitution error

City Centre is added
↓
insertion error

Updated dataset for E2E

updated versions

- cleaned (Dušek et al. 2019)
- enriched (Ferreira et al. 2021)

→ still contain a certain number of errors

Error Type	original E2E dataset			Cleaned dataset			Enriched dataset		
	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

E2E refined dataset

developed Python programs to convert the original E2E dataset to the E2E refined dataset

Text Refinement: Error Correction

indefinite articles

- For **an** child friendly, average coffee shop serving fast food try The Eagle, riverside near Burger King.

irregular MR values

- **Clear Hall** is known for Fast food and coffee shop style bakeries although, customers only rate them average the **Clowns** are quite amusing.

overlaps

- 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.

typos

- Moderately priced fast **found** can be found at Blue Spice in city centre.

Text Refinement: Error Correction

indefinite articles

- For **an** child friendly, average coffee shop serving fast food try The Eagle, riverside near Burger King.

a

irregular MR values

- **Clear Hall** is known for Fast food and coffee shop style bakeries although, customers only rate them average the **Clowns** are quite amusing.

overlaps

- 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.

typos

- Moderately priced fast **found** can be found at Blue Spice in city centre.

Text Refinement: Error Correction

- **indefinite articles**
  - For **an** child friendly, average coffee shop serving fast food try The Eagle, riverside near Burger King.

(removed)

- **irregular MR values**
  - **Clear Hall** is known for Fast food and coffee shop style bakeries although, customers only rate them average the **Clowns** are quite amusing.

- **overlaps**
  - 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.

- **typos**
  - Moderately priced fast **found** can be found at Blue Spice in city centre.

Text Refinement: Error Correction

- **indefinite articles**
  - For **an** child friendly, average coffee shop serving fast food try The Eagle, riverside near Burger King.
- **irregular MR values**
  - **Clear Hall** is known for Fast food and coffee shop style bakeries although, customers only rate them average the **Clowns** are quite amusing.
- **overlaps**
  - 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.

(removed)

- **typos**
  - Moderately priced fast **found** can be found at Blue Spice in city centre.

Text Refinement: Error Correction

- **indefinite articles**
  - For **an** child friendly, average coffee shop serving fast food try The Eagle, riverside near Burger King.
- **irregular MR values**
  - **Clear Hall** is known for Fast food and coffee shop style bakeries although, customers only rate them average the **Clowns** are quite amusing.
- **overlaps**
  - 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.
- **typos**
  - Moderately priced fast **found** can be found at Blue Spice in city centre.

food

Text Refinement: Normalization

British English

- Cheap family **favorite**, The Twenty Two, near The Rice Boat in riverside, got a 5 out of 5.

capital letters

- **a** coffee shop Zizzi located by the riverside has a high price range with an average customer rating. **t**hey are children friendly.

currency expression / prices

- The Punter is a Japanese restaurant under **20 pounds**.

quotation marks

- A highly rated coffee shop **``**The Punter**”** serving English food priced between £20 - £25 and is child friendly.

Text Refinement: Normalization

British English

favourite

- Cheap family **favorite**, The Twenty Two, near The Rice Boat in riverside, got a 5 out of 5.

capital letters

- **a** coffee shop Zizzi located by the riverside has a high price range with an average customer rating. **t**hey are children friendly.

currency expression / prices

- The Punter is a Japanese restaurant under **20 pounds**.

quotation marks

- A highly rated coffee shop ``The Punter'' serving English food priced between £20 - £25 and is child friendly.

Text Refinement: Normalization

British English

- Cheap family **favorite**, The Twenty Two, near The Rice Boat in riverside, got a 5 out of 5.

capital letters

- **a** coffee shop Zizzi located by the riverside has a high price range with an average customer rating. **t**hey are children friendly.

A

They

currency expression / prices

- The Punter is a Japanese restaurant under **20 pounds**.

quotation marks

- A highly rated coffee shop ``The Punter'' serving English food priced between £20 - £25 and is child friendly.

Text Refinement: Normalization

- **British English**
  - Cheap family **favorite**, The Twenty Two, near The Rice Boat in riverside, got a 5 out of 5.
- **capital letters**
  - **a** coffee shop Zizzi located by the riverside has a high price range with an average customer rating. **t**hey are children friendly.
- **currency expression / prices**
  - The Punter is a Japanese restaurant under **20 pounds**.

£20

- **quotation marks**
  - A highly rated coffee shop ``The Punter'' serving English food priced between £20 - £25 and is child friendly.

Text Refinement: Normalization

- **British English**
  - Cheap family **favorite**, The Twenty Two, near The Rice Boat in riverside, got a 5 out of 5.
- **capital letters**
  - **a** coffee shop Zizzi located by the riverside has a high price range with an average customer rating. **t**hey are children friendly.
- **currency expression / prices**
  - The Punter is a Japanese restaurant under **20 pounds**.
- **quotation marks**
  - A highly rated coffee shop ``The Punter'' serving English food priced between £20 - £25 and is child friendly.

MR Refinement

- **re-annotate all MR data manually**
  - developed Python programs for annotation
- **re-define all variations of MR values**

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

newly added

replaced from "high"

MR Refinement (cont.)

additional annotations

- MR order
- number of sentences
- sentence indexes

	attribute	value	order	sentence index
MR	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
Text	Moderately priced THE OLIVE GROVE pub is located on the riverside. It welcomes kids.			
number of sentences	2			

MR-Text Pair Refinement

delexicalization

- The value for “name” and “near” attributes are always directly reflected in the sentences

→ substitute the value in MR and sentence with special strings, “NAME” and “NEAR”

	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
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.			
number of sentences	2			

Experiments

- **method**
  - TGen (Dušek et al. 2016)
    - LSTM-based sequence-to-sequence model with an attention mechanism
    - used as the baseline method for the E2E NLG Challenge
  - trained 2 models
    - (A) training dataset in the E2E dataset
    - (B) training dataset in the E2E refined dataset
- **metrics**
  - BLEU, NIST, METEOR, ROUGE_L, CIDEr
    - used for the E2E NLG Challenge
- **results**
  - the performance of the model (B) outperformed that of the model (A)

model	training data	BLEU(↑)	NIST(↑)	METEOR(↑)	ROUGE_L(↑)	CIDEr(↑)
(A)	E2E dataset	0.5462	7.6209	0.4103	0.6561	2.2448
(B)	E2E Refined Dataset	0.5581	7.8378	0.4252	0.6488	2.3865

accurate label information led to improved performance

Summary

E2E Refined Dataset

- refined version of E2E dataset (MR-to-Text)
- reduced deletion / insertion / substitution errors in the original E2E dataset
  - text refinement
    - error correction
    - normalization
  - MR refinement
    - labelling annotations manually
  - MR-text pair refinement
    - delexicalization
- additional annotation
  - MR order
  - number of sentences
  - sentence indexes

Python programs are available