Constrained Decoding for Neural NLG from Compositional Representations

Anusha Balakrishnan, Jinfeng Rao, Kartikeya Upasani, Michael White, Rajen Subba

Facebook AI
Outline

• Demonstrate motivation for **compositional inputs** to NLG systems
• Introduce **constrained decoding** approach that improves semantic correctness of neural NLG system
• Put the two together for increased control, expressiveness, and correctness
Background
E2E NLG Dataset

~50K (meaning representation, sentence) pairs for the restaurants domain

**MR:** name[JJs’ Pub] rating[5 out of 5] familyFriendly[no] eatType[restaurant] near[Crowne Plaza Hotel]

**Sentence:** JJ’s pub is not a family friendly restaurant. It has a high customer rating of 5 out of 5. You can find it near the Crowne Plaza Hotel.
E2E Dataset and Shortcomings

- Provided crowdworkers with MR and asked them to write responses
- Lots of diversity in dataset, e.g. argument grouping, contrast, rich language
  - [But hard to control these aspects]
- **However**, models trained on the data lack this diversity
  - e.g. contrast only occurs 0.4% of the time in model generations

Possible reason: **Same MR → different discourse structures**
JJ’s pub is not a family friendly restaurant. It has a high customer rating of 5 out of 5. You can find it near the Crowne Plaza Hotel.

JJ’s Pub is not family friendly, but it has a high customer rating of 5 out of 5. It is a restaurant near the Crowne Plaza Hotel.
Reed et al. (2018)

- Proposed adding tokens to indicate target discourse structure (contrast/no contrast, # of sentences)
- Greatly improved accuracy of contrast expression and # sentences
- **But**: no way to control *which slots* gets contrasted!

*How can we control the expression of discourse relations?*
Why control discourse?

Prior work has shown increases in **perceived naturalness** when user models or preferences influence discourse.

What can you tell me about JJ’s Pub? I’d like to eat English food.

JJ’s pub is highly rated, *but it serves English food.*

5/5 rating and English food are usually contrasted.

Discourse-unaware system

(Lemon et al. 2004; Carenin and Moore, 2006; Walker et al. 2007; White et al. 2010; Demberg et al. 2011)
Why control discourse?

Prior work has shown increases in **perceived naturalness** when user models influence discourse.

What can you tell me about JJ’s Pub? I’d like to eat English food.

JJ’s Pub is highly rated, and it serves English food.

Discourse-aware system

(Lemon et al. 2004; Carenin and Moore, 2006; Walker et al. 2007; White et al. 2010; Demberg et al. 2011)
Compositional Meaning Representations
Proposed Representation Components

- **Arguments** e.g. `date_time`, `family_friendly`
  - Entities or slots to be mentioned
  - Can be nested (e.g. `date_time` contains `week_day`)
- **Dialog acts** e.g. `INFORM`, `RECOMMEND`
  - Contain arguments
- **Discourse relations** e.g. `JUSTIFY`, `CONTRAST`
  - Relationship between dialog acts / discourse relations

(Mann and Thompson, 1988; Rambow et al. 2001; Reiter and Dale, 2000; Walker et al. 2007)
Putting it all together

- Tree-structured representation
- Allows arbitrary nesting of relations and acts

```
[JOIN
 [CONTRAST
  [INFORM [name JJs’ Pub] [rating 5 out of 5] ]
  [INFORM [familyFriendly no ] ]
]
[INFORM [eatType restaurant] [near Crowne Plaza Hotel] ]
```
Data
Dataset Creation

User query crowdsourcing → Compositional MR generation → Response generation and annotation → Response quality evaluation → 33K examples for weather domain

Rule-based

https://github.com/facebookresearch/TreeNLG
Dataset Creation

User query crowdsourcing -> Compositional MR generation -> Response generation and annotation -> Response quality evaluation

Rule-based

Date: October 5
Location: Austin

How heavy is the rain expected to be?

INFORM[
  condition_not[rain]
  date_time[ month[October] day[5] ]
  location[ city[Austin] ]
]

INFORM[
  cloud_coverage[partly cloudy]
  temp_high[50] temp_low[37]
  date_time[ month[October] day[5] ]
  location[ city[Austin] ]
]

No rain is expected today in Austin. It’ll be partly cloudy with a high of 50 and a low of 37.

INFORM No [CONDITION_NOT rain] is expected [DATE_TIME [COLLOQUIAL today] in [LOCATION [CITY Austin] ]]. [INFORM It’ll be [CLOUD_COVERAGE partly cloudy] with a high of [TEMP_HIGH 50] and a low of [TEMP_LOW 37].]
Modified E2E Dataset

- Used the Berkeley neural parser to automatically infer CONTRAST and JOIN relations in the E2E challenge dataset
- Used Slug2Slug token tagger and HarvardNLP latent segmentation model to annotate responses
Approach
Model overview

Standard Seq2Seq model with attention

Input: Linearized tree structure

[CONTRAST [INFORM [condition rain ] [date_time [week_day Saturday ] ] ] [INFORM [cloud_coverage sunny ] [date_time [week_day Sunday ] ] ] ] ]

Output: Response with tree structure preserved

[CONTRAST [INFORM There’ll be [condition rain ] on [date_time [week_day Saturday ] ] ], but [INFORM [date_time [week_day Sunday ] ] will be [cloud_coverage sunny ] ] ].
Semantic Correspondence

Leverage correspondence between input and output structures

[CONTRAST [INFORM [condition rain] [date_time [week_day Saturday]]] [INFORM [cloud_coverage sunny] [date_time [week_day Sunday]]]]

[JOIN [INFORM There’ll be [condition rain] on [date_time [week_day Saturday]]], and [INFORM [date_time [week_day Sunday]] will be [cloud_coverage sunny]].]
Semantic Correspondence

Leverage correspondence between input and output structures

CONTRAST
INFORM condition rain ] [date_time [week_day Saturday ] ]
INFORM cloud_coverage sunny ] [date_time [week_day Sunday ] ]

JOIN
INFORM There’ll be [condition rain ] on [date_time [week_day Saturday ] ]], and [INFORM [date_time [week_day Sunday ] ] will be [cloud_coverage sunny ] ]].

Tree accuracy metric for semantic correctness
Constrained Decoding

Idea: During beam search, invalidate tokens that would result in responses that don’t match the input structure (based on the nonterminals in the output)

\[
\text{[INFORM [condition_not rain] [date_time [colloquial today ] ] ]}
\]
Experiments and Results
Models

- **Flat**
  - Input and output contain a flat list of arguments

- **Token**
  - Inspired by Reed et al. 2018, input is similar to Flat but has tokens indicating # of Contrast and Join relations

- **Tree**
  - Input and output are linearized tree representations

- **Constr**
  - Tree with constrained decoding
Metrics

- Automatic
  - BLEU
  - Tree accuracy
- Human evaluation
  - Grammaticality
  - Semantic correctness
  - Disc. correctness (semantic correctness measured on challenging subset)
Results

Tree-based representations improve correctness
Results

Constrained decoding improves correctness further
Observations

- Constrained decoding can still result in incorrect generations
  - Model generates surface forms without generating the non-terminal
  - [restaurant <name>] is not family friendly and serves [cuisine Indian] food.
- Occasional (<1%) stuttering due to constrained decoding
- Unnatural phrasings
  - Yes, you should wear [ATTIRE an umbrella]
Conclusions

- Shown that tree structured representations can greatly improve controllability of generated text
- Introduced a simple constrained decoding technique that is only applied to the decoder
- Released a conversational NLG corpus for the weather domain

Check out our dataset and code:

https://github.com/facebookresearch/TreeNLG
Thank you!

https://github.com/facebookresearch/TreeNLG
Future Work

- Condition on the user query for increased naturalness in context
- Improving grammaticality and naturalness of generated text
- Checking outputs with reverse models
- Using constraints in training
Modified E2E Dataset

- Used the Berkeley neural parser to automatically infer CONTRAST and JOIN relations in the E2E challenge dataset
- Used Slug2Slug token tagger and HarvardNLP latent segmentation model to annotate responses

```
[JOIN [CONTRAST [INFORM [NAME JJ’s Pub] is not [FAMILY_FRIENDLY_NO family friendly ]], [INFORM but has a [RATING_5_OUT_OF_5 high customer rating of 5 out of 5]]]. [INFORM It is a [EATTYPE_RESTAURANT restaurant] near the [NEAR Crowne Plaza Hotel].]]

[JOIN [CONTRAST [INFORM [name JJs’ Pub] [rating 5 out of 5]] [INFORM [familyFriendly no]]]]

[INFORM [eatType restaurant] [near Crowne Plaza Hotel]]
```
Constrained Decoding (contd.)

- Invalidate tokens that would result in responses that don’t match the input structure (based on the nonterminals in the output)
  - Don’t allow opening non-terminals (e.g. [CONTRAST ) that aren’t allowed in that part of the subtree
  - Don’t allow closing non-terminals ( ] ) if subtree isn’t complete
    - Account for ellipsis/aggregation (if a value has already been expressed previously)
E2E Results

![E2E Dataset Graph](chart.png)
Results (contd.)

- Grammaticality is slightly lower as a result of constrained decoding, but not significantly.
- Tree-based models improve BLEU and diversity metrics compared to models based on flat MRs, like Slug2Slug which won the E2E challenge.
  - Diversity improves without negative impact on automatic metrics and/or semantic correctness!