Natural Language Generation from Structured Inputs

Sebastian Gehrmann, Falcon Dai, Henry Elder, Alexander Rush

OpenNMT Workshop Paris, February 2018

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Meaning Representation	name[The Golden Palace], eatType[coffee shop], food[Fast food], priceRange[cheap], customer rating[5 out of 5], area[riverside]
Human Reference	A coffee shop located on the riverside called The Golden Palace, has a 5 out of 5 customer rating. Its price range are fairly cheap for its excellent Fast food.

Data and Challenge by Novikova et al., 2017

Possible Attributes in the E2E NLG dataset

Attribute	Example
area	city centre, riverside, . . .
customerRating	1 out of 5, average, . . .
eatType	coffee shop, restaurant, . . .
familyFriendly	yes / no
food	Chinese, English, . . .
name	Wildwood, The Wrestlers, . . .
near	Café Sicilia, Clare Hall, . . .
priceRange	less than £20, cheap, . . .

Stages of Natural Language Generation

In the past, the task has been broken into the stages

- Content Planning $\leftarrow$ done for us

  Select information for the generation.

- Sentence Planning

  Choose words and structures to fit information into sentences.

- Surface Realization

  Choose syntax, morphology, and orthography for natural text.

One Input, Multiple Texts

Input

Name: *Alimentum*

Area: *Riverside*

Family-Friendly: *no*

Near: *Burger King*

Output

Located off the river near Burger King, Alimentum does not allow families.

Alimentum is a non-family-friendly establishment near Burger King at the riverside.

Alimentum is not family-friendly. It is located near Burger King in riverside.

Alimentum across from Burger King no kids

Input

Name: *Alimentum*

Area: *Riverside*

Family-Friendly: *no*

Near: *Burger King*

Output

Located off the river near Burger King, Alimentum does not allow families.

Alimentum is a non-family-friendly establishment near Burger King at the riverside.

Alimentum is not family-friendly. It is located near Burger King in riverside.

Alimentum across from Burger King no kids

One Input, Multiple Texts

Input

Name: *Alimentum*

Area: *Riverside*

Family-Friendly: *no*

Near: *Burger King*

Output

Located off the river near Burger King, Alimentum does not allow families.

Alimentum is a non-family-friendly establishment near Burger King at the riverside.

Alimentum is not family-friendly. It is located near Burger King in riverside.

Alimentum across from Burger King no kids

E2E NLG as S2S Problem

To generate an input sequence, we introduce start and end tokens:

area[city centre]

$\rightarrow$

__start_area__ city centre __end_area__

Names are not delexicalized and attributes appear in the same order.

Let $(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}), \ldots (\mathbf{x}^{(N)}, \mathbf{y}^{(N)}) \in (\mathcal{X}, \mathcal{Y})$ be the N source/target pairs.

We learn a function f parametrized by θ that maximizes the conditional probability of $p_\theta(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{n} p_\theta(y_t|\mathbf{y}_{[t-1]}, \mathbf{x})$.

E2E NLG as S2S Problem

To generate an input sequence, we introduce start and end tokens:

area[city centre]

$$\rightarrow$$

__start_area__ city centre __end_area__

Names are not delexicalized and attributes appear in the same order.

Let $(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}), \ldots (\mathbf{x}^{(N)}, \mathbf{y}^{(N)}) \in (\mathcal{X}, \mathcal{Y})$ be the N source/target pairs.

We learn a function f parametrized by θ that maximizes the conditional probability of $p_\theta(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{n} p_\theta(y_t|\mathbf{y}_{[t-1]}, \mathbf{x})$.

Problem Specification

Encoder and Decoder are RNNs (or Transformer):

$$\mathbf{h}^x_m \leftarrow \mathrm{RNN}(\mathbf{h}^x_{m-1}, x_m) \quad \mathbf{h}_t \leftarrow \mathrm{RNN}(\mathbf{h}_{t-1}, w_t)$$

Attention:

$$p_{att}(t) \leftarrow \mathrm{softmax}([\mathbf{h}^x_1; \dots; \mathbf{h}^x_M]^\top \mathbf{h}_t)$$

$$\mathbf{c}_t \leftarrow \mathbb{E}_{m \sim p_{att}}[\mathbf{h}^x_m] = \sum_{m=1}^{M} p_{att}(m)\mathbf{h}^x_m$$

$$\mathbf{o}_t = \tanh([\mathbf{c}_t, \mathbf{h}_t]\mathbf{W}_{out} + b_{out})$$

Word Generation:

$$p_{vocab} = \mathrm{softmax}(\mathbf{o}_t \mathbf{W}_{gen} + b_{gen})$$

Words are often copied into the target

name: **The Eagle** food: **Italian** customerRating: **1 out of 5**

The Eagle is an **Italian** joint with a customer rating of **1 out of 5** .

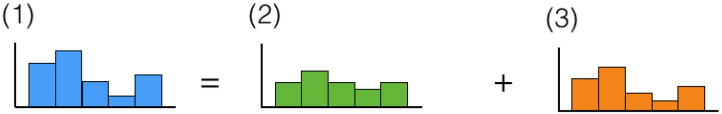
Copy-Mechanism

Gülcehre et al. (2016), See et al. (2017)

- We introduce a binary variable z_t for each decoding step t that acts as switch between copying and generating a word.

$$p(y_t, z_t | y_{[t-1]}, \mathbf{x}) = \sum_{z \in \{0,1\}} p(y_t, z_t = z | \mathbf{y}_{[t-1]}, \mathbf{x})$$

- z_t is computed such $p(z_t) = \sigma(\mathbf{o}_t^T v)$. Then, the joint probability is decomposed into the two terms



$$p(y_t | y_{[t-1]}, \boldsymbol{x}) = p(z_t=1)\ p(y_t | z_t=1) + p(z_t=0)\ p(y_t | z_t=0)$$

Compute copy distribution

$$p_{copy}(m) \leftarrow \mathrm{softmax}([\mathbf{h}_1^x; \ldots; \mathbf{h}_M^x]^\top \mathbf{o}_n)$$

Prediction

$$\begin{aligned} p_{gen} &= \sigma(\mathbf{o}_n^T v) \\ p(w_{n+1} | w_{1:n}, x_{1:M}) &= p_{gen} \times p_{vocab} \\ &\quad + (1 - p_{gen}) \times \mathbb{E}_{m \sim p_{copy}}[\mathbf{1}(w_{n+1} = x_m)] \end{aligned}$$

The Learned Model is Missing Attributes

Input	**Output**
Name: *The Mill*	
eat type: *restaurant*	
food: *English*	
price range: *less than £20*	The Mill is a low-priced restaurant near Café Rouge.
customer rating: *low*	
area: *riverside*	
family-friendly: *no*	
near: *Café Rouge*	

Attributes in red are not part of the generated text.

Coverage and Length Penalty

Wu et al. (2016)

- To ensure that an equal amount of attention is given to every input, we penalize words that receive a total attention over 1.0.

$$\text{coveragePenalty}(\mathbf{x}, \mathbf{y}) = \beta \cdot \sum_{i=1}^{|\mathbf{x}|} \log(\min(\sum_{t=1}^{|\mathbf{y}|} a_i^t, 1.0)).$$

- Scores are normalized by length so that short sentences are not preferred over longer ones. The loss is divided by

$$\text{lengthPenalty}(\mathbf{y}) = \frac{(5 + |\mathbf{y}|)^{\alpha}}{(5 + 1)^{\alpha}}$$

Another approach by Tu et al. (2016) builds this mechanism into the attention.

Template Hallucination

Input

Name: *The Vaults*

eat type: *pub*

price range: *high*

customer rating: *high*

family-friendly: *yes*

near: *Rainbow Vegetarian Café*

Output

The Vaults is an expensive, three star, family friendly pub located near the Rainbow Vegetarian Café.

Oops...

Template Hallucination

Input

Name: *The Vaults*

eat type: *pub*

price range: *high*

customer rating: *high*

family-friendly: *yes*

near: *Rainbow Vegetarian Café*

Output

The Vaults is an expensive, three star, family friendly pub located near the Rainbow Vegetarian Café.

The phrase *three star* occurs 271 times in the training set (0.6%), always in the context of *average* ratings.

The phrase *expensive, three star* occurs four times.

Diverse Ensembling

Guzman-Rivera et al. (2012), Lee et al. (2016)

- Train separate models $f_1, \ldots, f_K$.
- Each data point is assigned to one model.
- Jointly optimize the
  - Assignments of data points to models
  - Parameters of each model
- Let $w \sim \text{Cat}(1/K)$ be the weights of the models.
- The overall objective for the joint optimization becomes

$$\text{argmin}_{w,\theta} \sum_{i=1}^{|\mathcal{X}|} \sum_{k=1}^{K} w_k^i \cdot \mathcal{L}(\mathbf{y}^i, f_k(\mathbf{x}^i)),$$

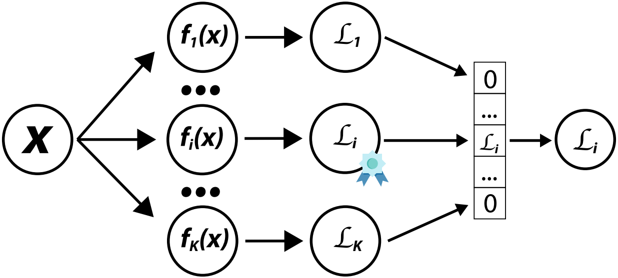
Diverse Ensembling

Guzman-Rivera et al. (2012), Lee et al. (2016)

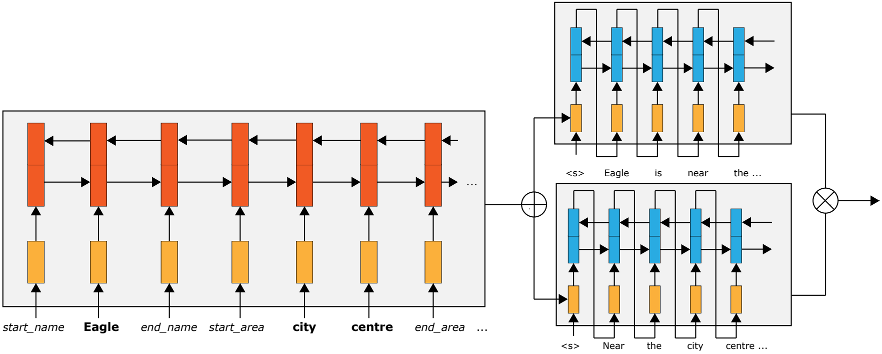
- Train separate models $f_1, \ldots, f_K$.
- Each data point is assigned to one model.
- Jointly optimize the
  - Assignments of data points to models
  - Parameters of each model
- Let $w \sim \mathsf{Cat}(1/K)$ be the weights of the models.
- The overall objective for the joint optimization becomes

$$\mathsf{argmin}_{w,\theta} \sum_{i=1}^{|\mathcal{X}|} \sum_{k=1}^{K} w_k^i \cdot \mathcal{L}(\mathbf{y}^i, f_k(\mathbf{x}^i)),$$

- Optimization can be achieved via hard EM
- E-Step: $\hat{k} = \text{argmax}_{k \in [K]} p_\theta(\mathbf{y}|\mathbf{x}, w = k)$ (find the best model)
- M-Step: $\text{argmax}_\theta \; p_\theta(\mathbf{y}|\mathbf{x}, w = \hat{k})$ (update the best model)



Diverse Ensembling with Two Decoders



We can share a subset of parameters, in this case the encoder.

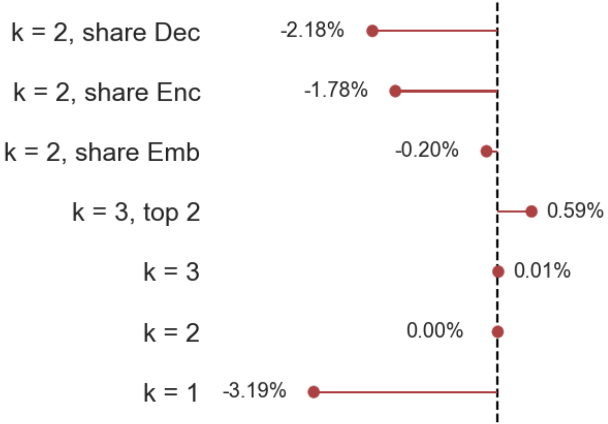
Results

A combination of the techniques lead to best ROUGE-L, CIDEr and METEOR scores among 70 submissions (BLEU 3rd, NIST 5th).

	BLEU	NIST	METEOR	ROUGE	CIDEr	Change
All	74.3	8.8	48.6	76.1	2.6	
- Cov/Length	73.2	8.6	48.1	75.6	2.5	-1.48%
- MCL Loss	69.8	8.2	47.8	74.3	2.5	-3.69%
- Copy	71.5	8.5	46.5	73.0	2.5	-3.51%

Investigation of Multiple Choice Loss

Relative Change from k = 2 while varying k and shared parameters



Conclusion

- S2S models with copy mechanism can effectively generate short text about a number of input attributes.
- Coverage and Length penalties lead to less ignored inputs.
- Multiple Choice Loss helps the model learn better latent sentence plans and makes the model more robust towards outliers in the training data.