

Controlling Text Generation

Alexander Rush

(with Yoon Kim, Sam Wiseman, Sebastian Gehrmann,
Yuntian Deng, Justin Chiu, and Demi Guo)



New York 2018

nlp.seas.harvard.edu / @harvardnlp

1 Preface: Seq2Seq+

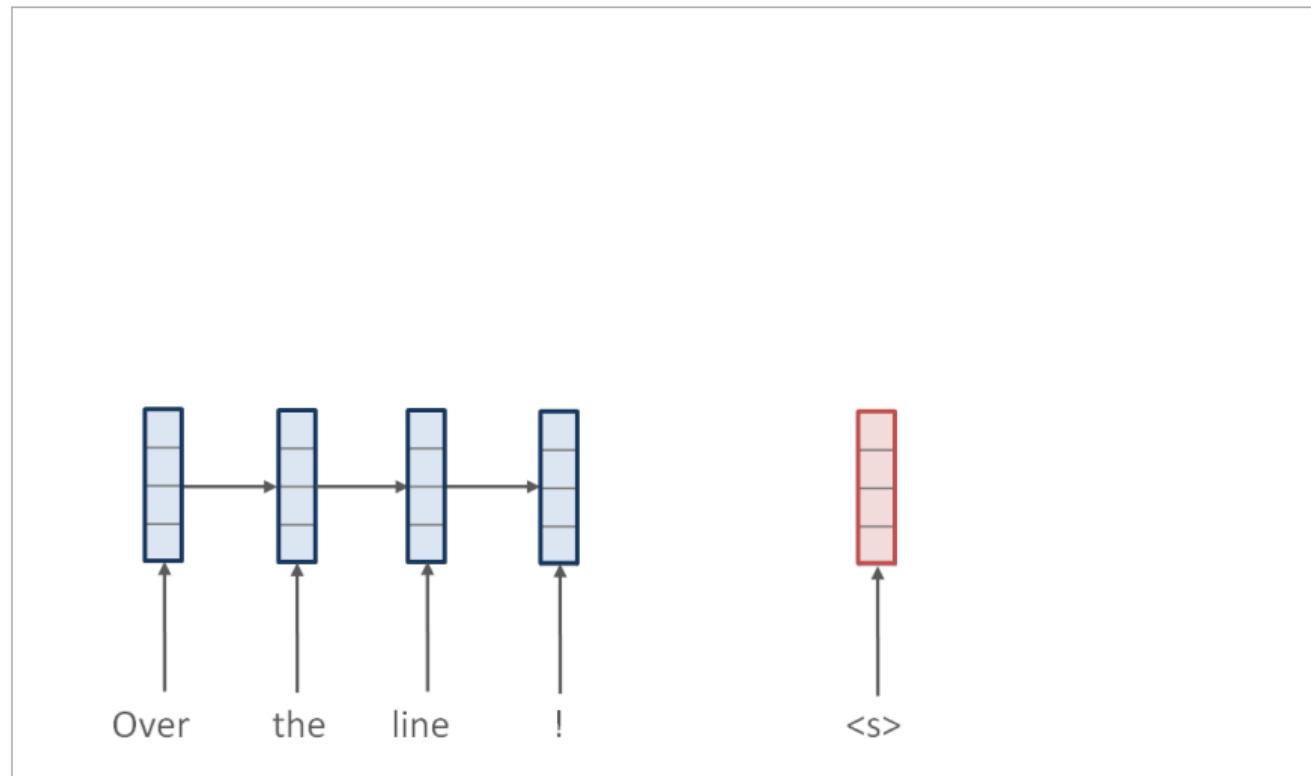
2 Introduction: Data-Driven Text Generation

3 Latent-Variable Generation

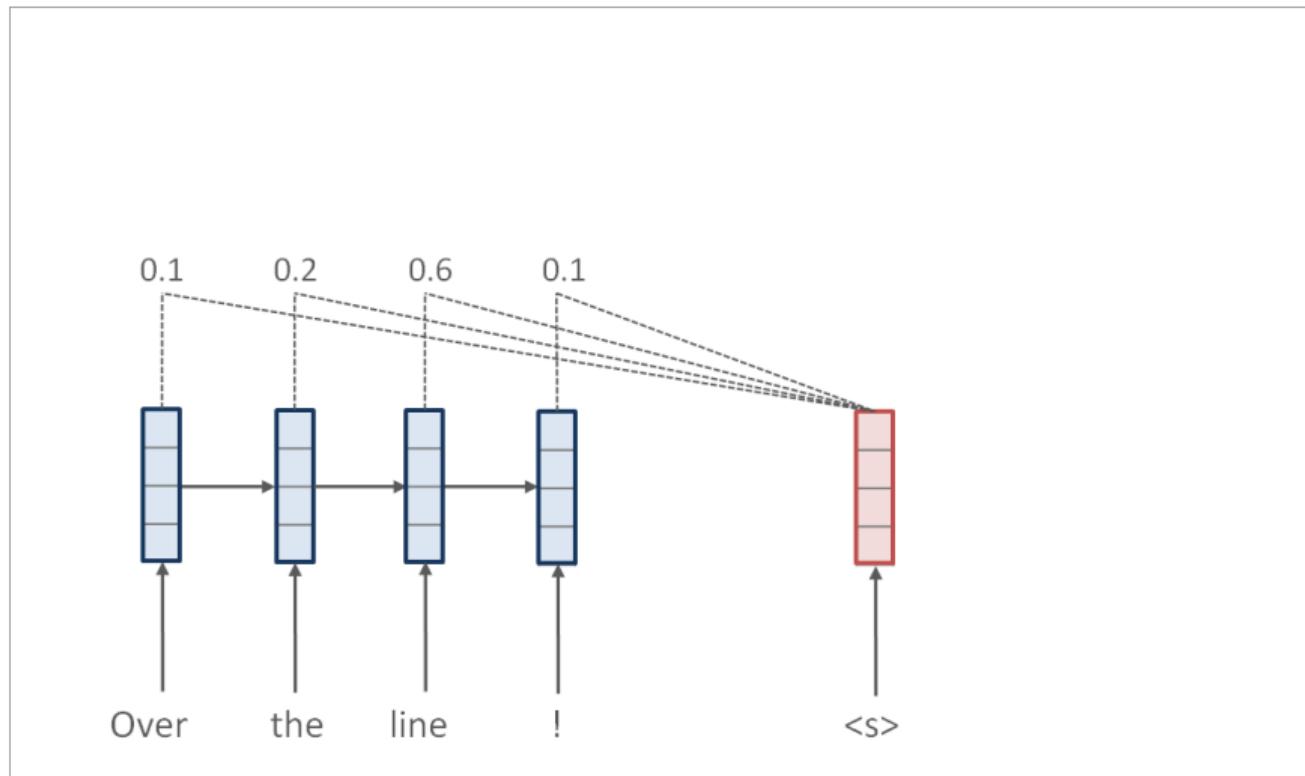
4 Work 1: Learning Neural Templates

5 Work 2: Learning Alignments

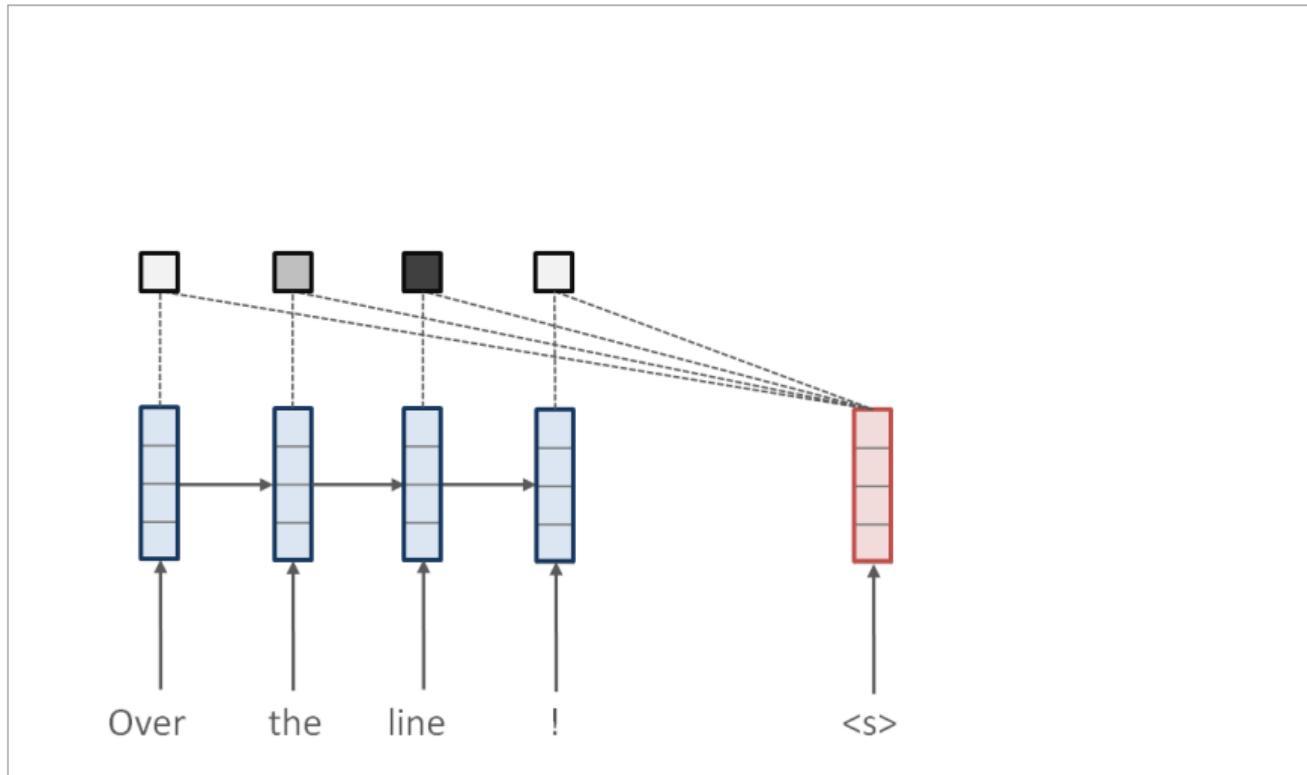
Seq2Seq+



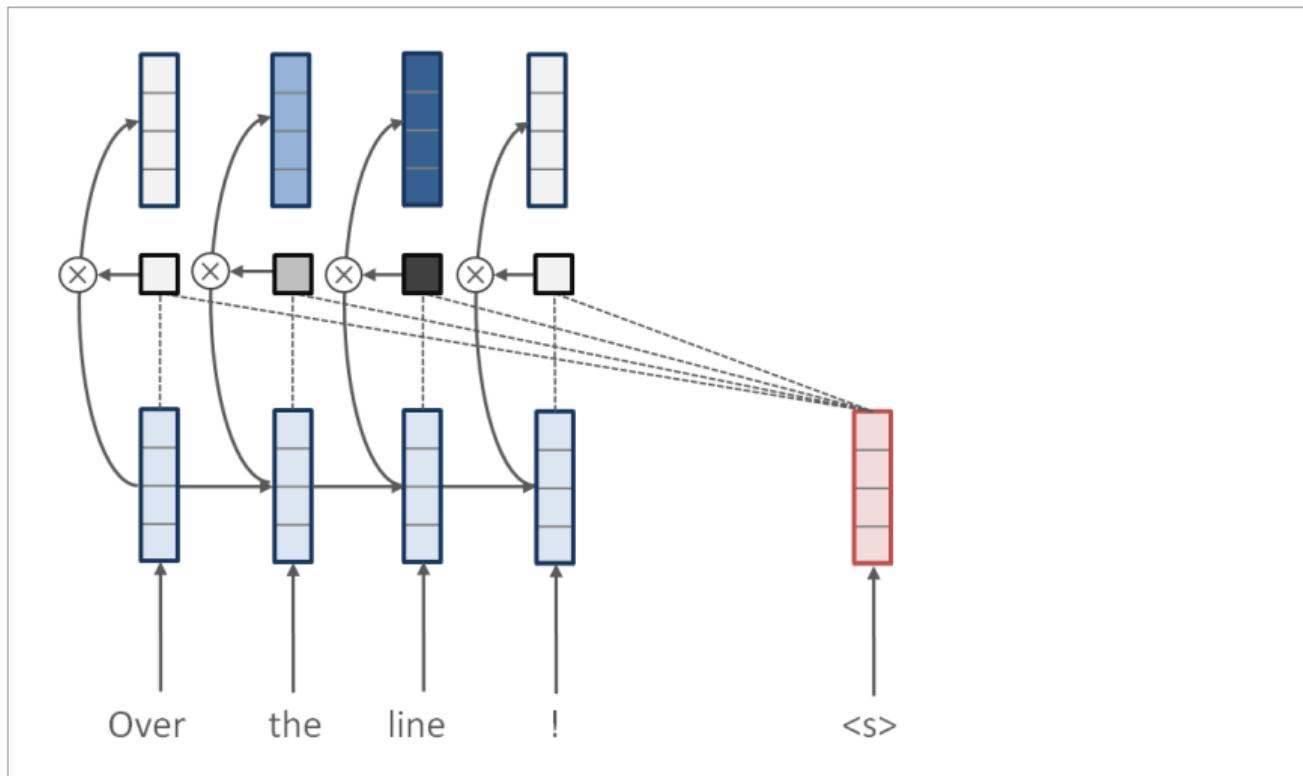
Seq2Seq+



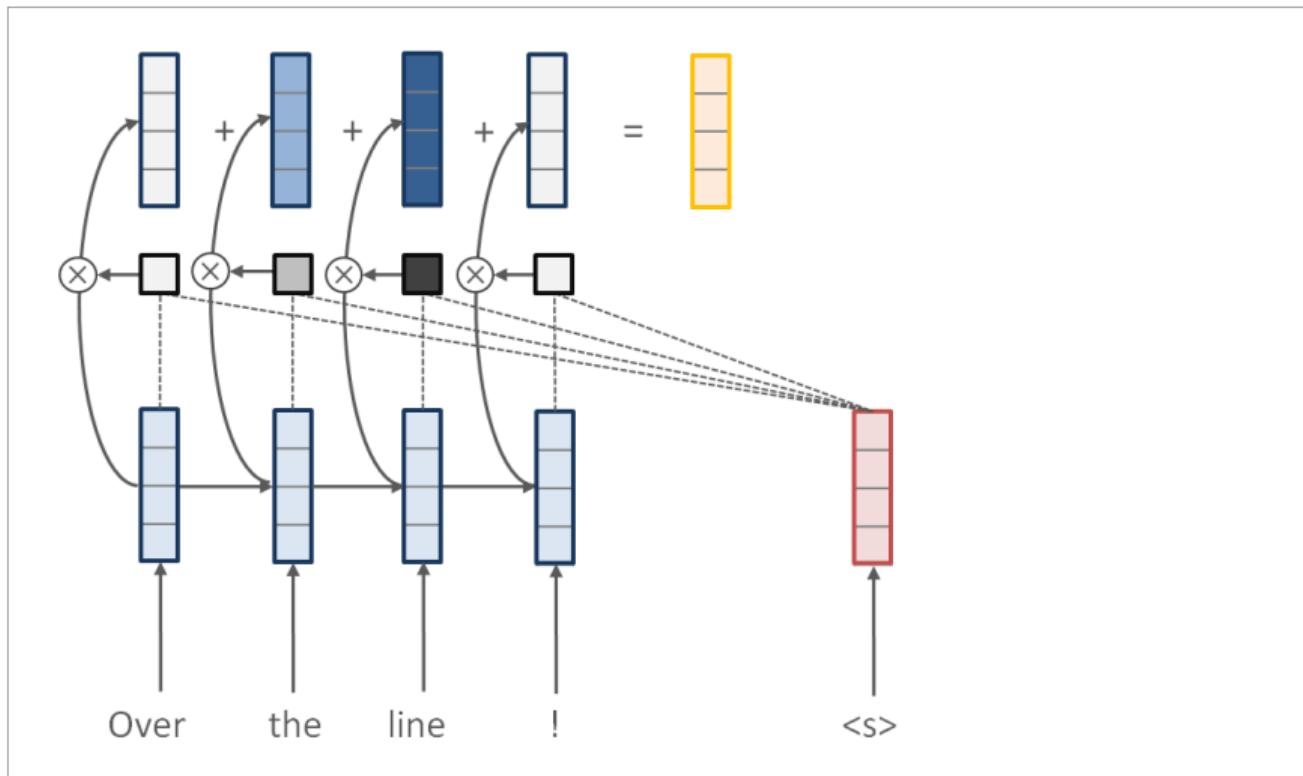
Seq2Seq+



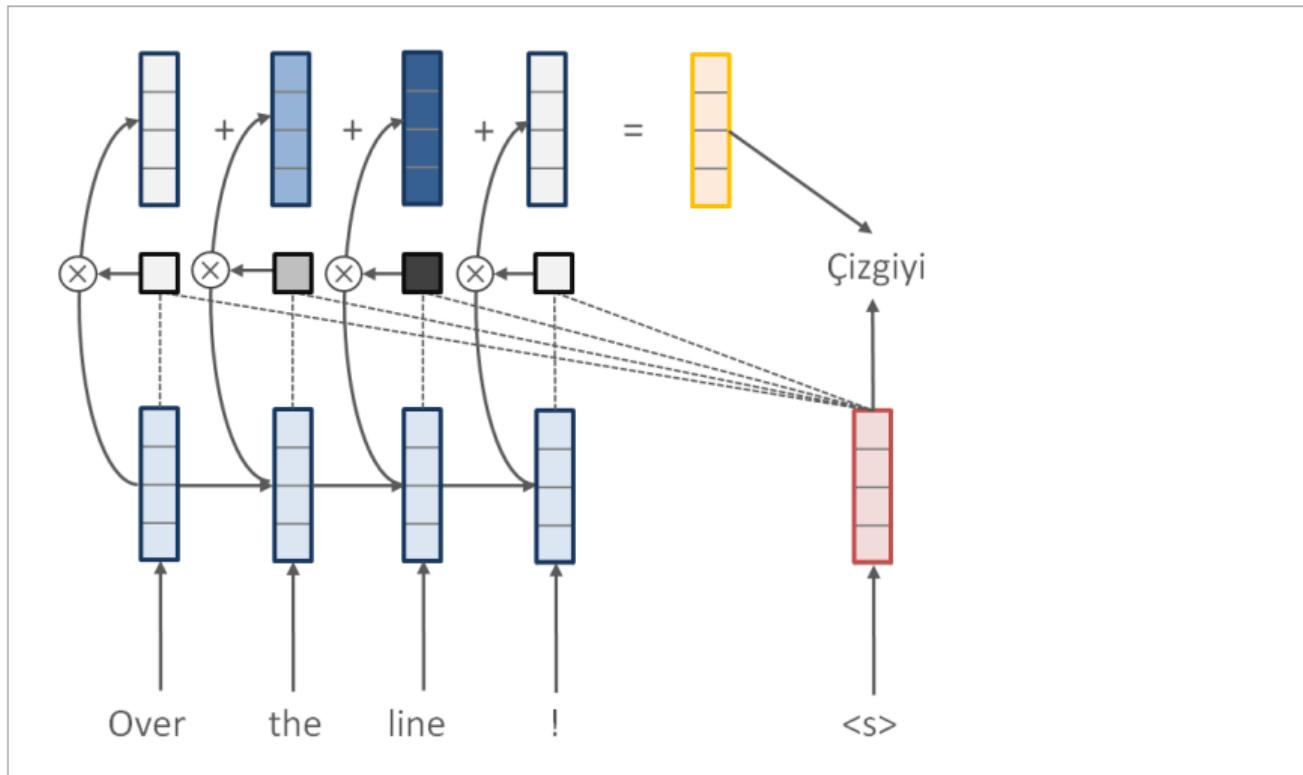
Seq2Seq+



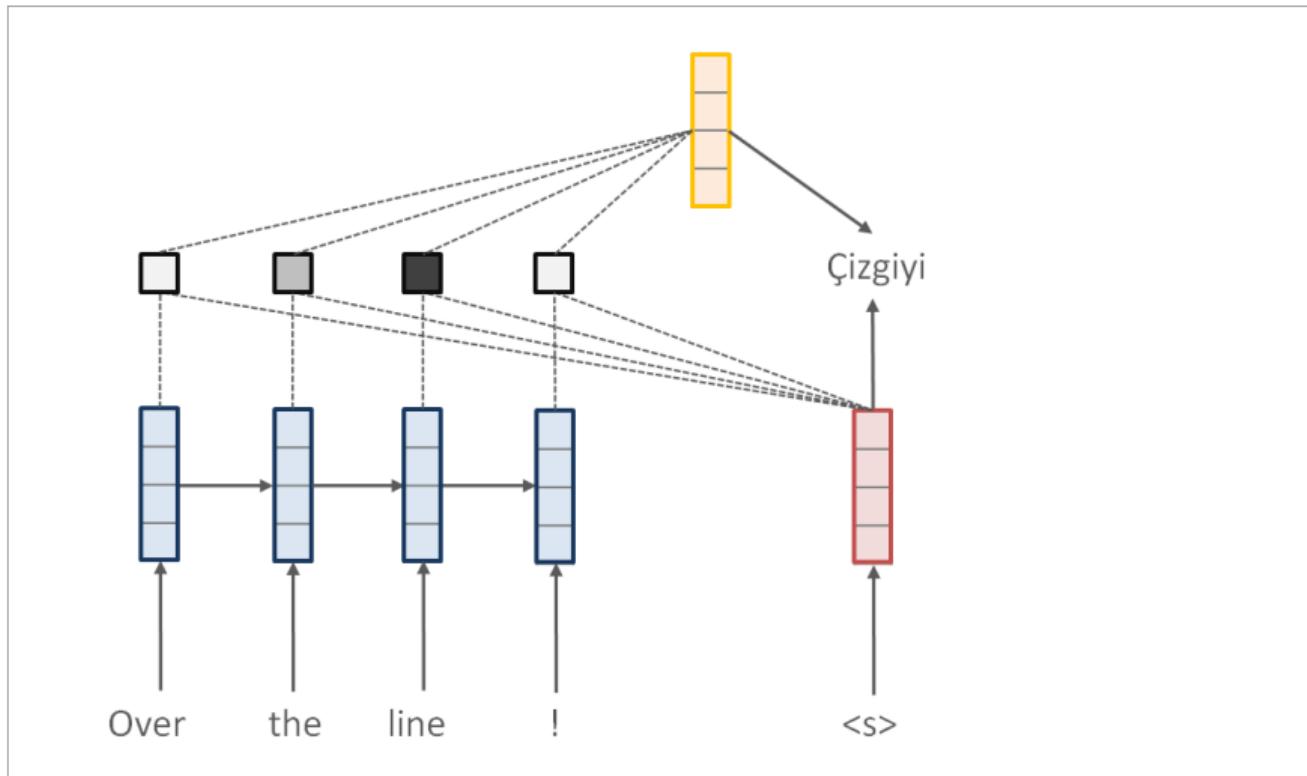
Seq2Seq+



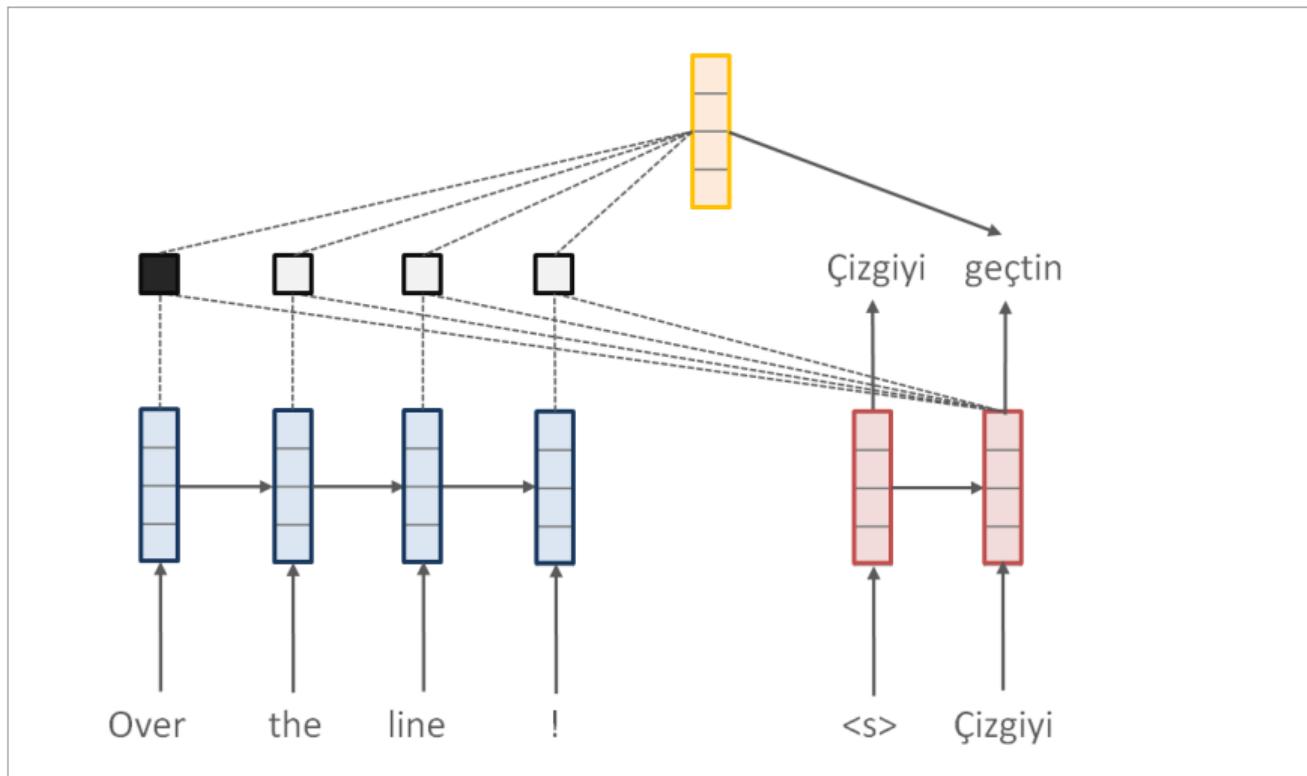
Seq2Seq+



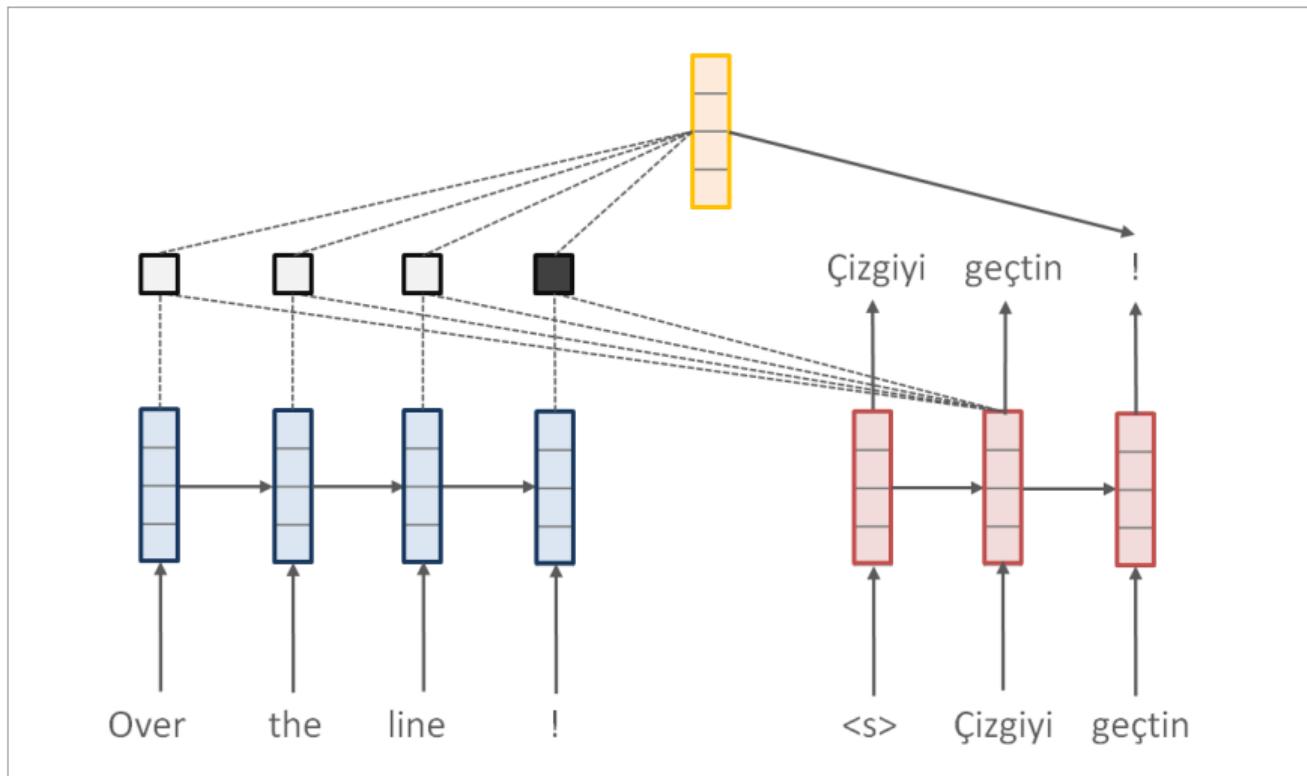
Seq2Seq+



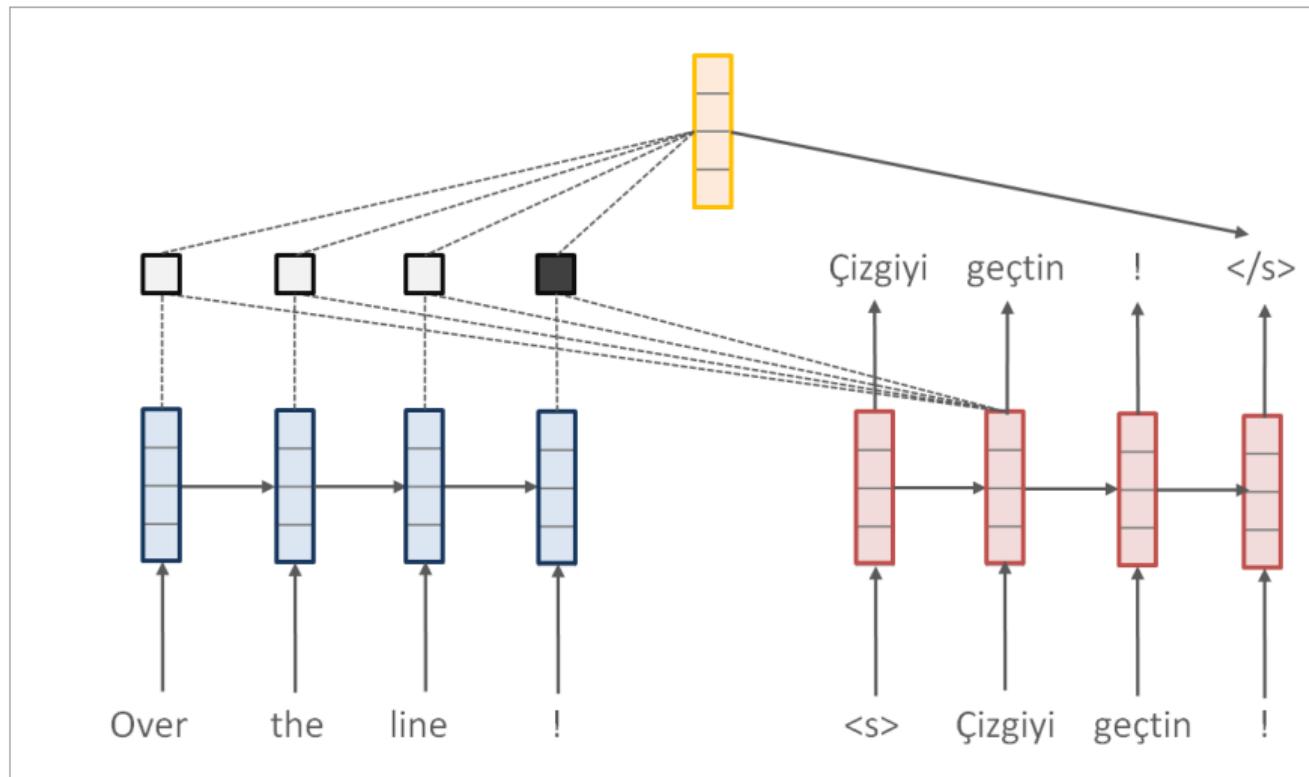
Seq2Seq+



Seq2Seq+



Seq2Seq+



Seq2Seq+

Encoder :

$$\mathbf{h}_m^x \leftarrow \text{RNN}(\mathbf{h}_{m-1}^x, x_m)$$

Attention

$$\alpha \leftarrow \text{softmax}([\mathbf{h}_1^x; \dots; \mathbf{h}_M^x]^\top \mathbf{h}_n)$$

$$\mathbf{c}_n \leftarrow \sum_{m=1}^M \alpha_m \mathbf{h}_m^x$$

Decoder:

$$\mathbf{h}_n \leftarrow \text{RNN}(\mathbf{h}_{n-1}, w_n)$$

Prediction

$$p(w_{n+1} | w_{1:n}, x_{1:M}) = \text{softmax}(\mathbf{W}[\mathbf{h}_n, \mathbf{c}_n])$$

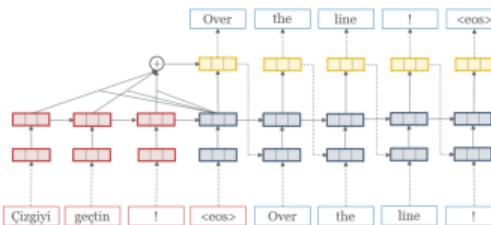
OpenNMT

(Klein et al, 2017)

The screenshot shows the OpenNMT website's home page. At the top is a large white logo 'NMT' with a stylized 'M'. Below it is a sub-headline: 'An open-source neural machine translation system.' A horizontal menu bar follows, containing links for English, Français, 简体中文, 한국어, 日本語, Русский, and العربية. The main navigation menu includes Home, Quickstart [Lua], Quickstart [Python], Advanced guide, Models and Recipes, FAQ, About, and Documentation.

Home

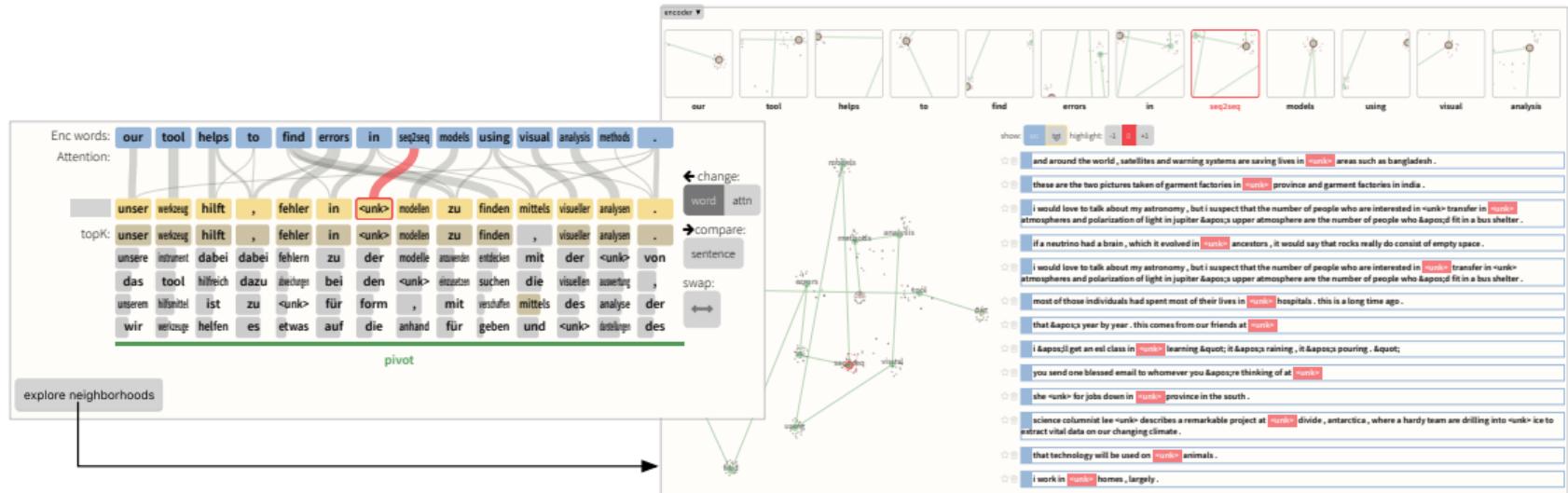
OpenNMT is a industrial-strength, open-source (MIT) neural machine translation system utilizing the Torch/PyTorch mathematical toolkit.



OpenNMT is used as provided in [production](#) by major translation providers. The system is designed to be simple to use and easy to extend, while maintaining efficiency and state-of-the-art translation accuracy.

Seq2Seq-Vis

(Strobelt et al, 2018)



1 Preface: Seq2Seq+

2 Introduction: Data-Driven Text Generation

3 Latent-Variable Generation

4 Work 1: Learning Neural Templates

5 Work 2: Learning Alignments

The Modern Text Generation Challenge



Text Generation: Talk about Text (Translation / Summarization)

mexico city , mexico -lrb- cnn -rrb- – heavy rains and flooding have forced hundreds of thousands of people from homes in southern mexico 's state of tabasco over the past four days , with nearly as many trapped by the rising waters , state officials said thursday . officials say about 300,000 people are still trapped by the worst flooding in the region for 50 years . the grijalva river pushed over its banks through the state capital of villahermosa on thursday , forcing government workers to evacuate and leaving up to 80 percent of the city flooded , gov. andres granier 's office told cnn . about 700,000 people have seen their homes flooded , with about 300,000 of those still trapped there , granier 's office reported . one death had been blamed on the floods , which followed weeks of heavy rain in the largely swampy state . tabasco borders guatemala to the south and the gulf of mexico to the north . . .



Text Generation: Talk about Text (Translation / Summarization)

mexico city , mexico -lrb- cnn -rrb- – heavy rains and flooding have forced hundreds of thousands of people from homes in southern mexico 's state of tabasco over the past four days , with nearly as many trapped by the rising waters , state officials said thursday . officials say about 300,000 people are still trapped by the worst flooding in the region for 50 years . the grijalva river pushed over its banks through the state capital of villahermosa on thursday , forcing government workers to evacuate and leaving up to 80 percent of the city flooded , gov. andres granier 's office told cnn . about 700,000 people have seen their homes flooded , with about 300,000 of those still trapped there , granier 's office reported . one death had been blamed on the floods , which followed weeks of heavy rain in the largely swampy state . tabasco borders guatemala to the south and the gulf of mexico to the north . . .



tabasco and chia-
pas states hardest
hit. authorities say
700,000 affected . . .

Text Generation:

Talk about Structured Data (Generation)

TEAM	W	L	PTS	...
Heat	11	12	103	...
Hawks	7	15	95	...



Text Generation: Talk about Structured Data (Generation)

TEAM	W	L	PTS	...
Heat	11	12	103	...
Hawks	7	15	95	...

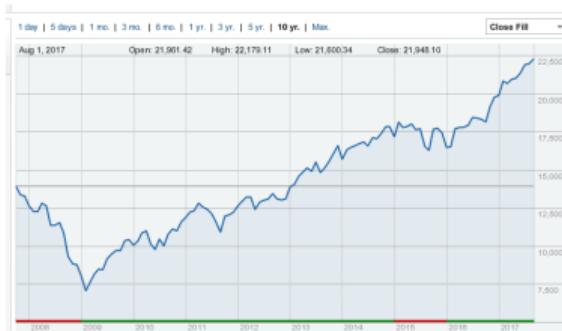


The Atlanta Hawks
defeated the Mi-
ami Heat, 103 - 95,
at Philips Arena on
Wednesday. Atlanta
...

Text Generation: Talk about the Environment (Multimodal)



Text Generation: Talk about the Environment (Multimodal)



Dow and S&P 500
close out week at
all-time highs ...

Are these problems just Seq2Seq?

- Certainly the key aspect of their success.
- However there has slowly tentative steps towards problem specific modeling.

Case Study: Seq2Seq+Copy

(Gulcehre et al, 2016) ...

Encoder :

$$\mathbf{h}_m^x \leftarrow \text{RNN}(\mathbf{h}_{m-1}^x, x_m)$$

Attention

$$\alpha \leftarrow \text{softmax}([\mathbf{h}_1^x; \dots; \mathbf{h}_M^x]^\top \mathbf{h}_n) \quad \mathbf{c}_n \leftarrow \sum_{m=1}^M \alpha_m \mathbf{h}_m^x$$

Decoder:

$$\mathbf{h}_n \leftarrow \text{RNN}(\mathbf{h}_{n-1}, w_n)$$

Prediction

$$\begin{aligned} p(w_{n+1}|w_{1:n}, x_{1:M}) = & \quad p(z=1|w_{1:n}, x_{1:M}) \times \text{softmax}(\mathbf{W}[\mathbf{h}_n, \mathbf{c}_n]) \\ & + \quad p(z=0|w_{1:n}, x_{1:M}) \times \sum_{m:x_m=w_{n+1}} \alpha_m \end{aligned}$$

Copy is Critical For Summarization

(See et al. 2017)

Pointer-generator model + coverage summary

```
francis saili has signed a two-year deal to join munster later this year .  
the 24-year-old was part of the new zealand under-20 side that won the junior world championship in italy in 2011 .  
saili 's signature is something of a coup for munster and head coach anthony foley .
```

Bottom-Up Summarization

(Gehrmann et al, 2018)

- Shows that controlling copying alone produces state-of-the-art results.

Copy is Critical For Summarization

(See et al. 2017)

Pointer-generator model + coverage summary

```
francis saili has signed a two-year deal to join munster later this year .  
the 24-year-old was part of the new zealand under-20 side that won the junior world championship in italy in 2011 .  
saili 's signature is something of a coup for munster and head coach anthony foley .
```

Bottom-Up Summarization

(Gehrmann et al, 2018)

- Shows that controlling copying alone produces state-of-the-art results.

Challenges of Neural Generation

(Wiseman et al, 2017)

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		
Dwight Howard	11	17	23	9		
Paul Millsap	2	9	21	8		
Goran Dragic	4	2	21	8		
Wayne Ellington	2	3	19	7		
Dennis Schroder	7	4	17	8		
Rodney McGruder	5	5	11	3		
...						

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 and forced them to commit 16 turnovers. Atlanta also dominated in the paint, winning the rebounding battle, 47 - 34, and outscoring them in the paint 58 - 26. The Hawks shot 49 percent from the field and assisted on 27 of their 43 made baskets. This was a near wire-to-wire win for the Hawks, as Miami held just one lead in the first five minutes. Miami (7 - 15) are as beat-up as anyone right now and it's taking a toll on the heavily used starters. Hassan Whiteside really struggled in this game, as

Challenges of Neural Generation

(Wiseman et al, 2017)

The Utah Jazz (38 - 26) defeated the Houston Rockets (38 - 26) 117 - 91 on Wednesday at Energy Solutions Arena in Salt Lake City . The Jazz got out to a quick start in this one , out - scoring the Rockets 31 - 15 in the first quarter alone . Along with the quick start , the Rockets were the superior shooters in this game , going 54 percent from the field and 43 percent from the three - point line , while the Jazz went 38 percent from the floor and a meager 19 percent from deep . The Rockets were able to out - rebound the Rockets 49 - 49 , giving them just enough of an advantage to secure the victory in front of their home crowd . The Jazz were led by the duo of Derrick Favors and James Harden . Favors went 2 - for - 6 from the field and 0 - for - 1 from the three - point line to score a game - high of 15 points , while also adding four rebounds and four assists

Challenges of Neural Generation (Wiseman et al, 2017)

The Utah Jazz (38 - 26) defeated the Houston Rockets (38 - 26) 117 - 91 on Wednesday at Energy Solutions Arena in Salt Lake City . The Jazz got out to a quick start in this one , out - scoring the Rockets 31 - 15 in the first quarter alone . Along with the quick start , the Rockets were the superior shooters in this game , going 54 percent from the field and 43 percent from the three - point line , while the Jazz went 38 percent from the floor and a meager 19 percent from deep . The Rockets were able to out - rebound the Rockets 49 - 49 , giving them just enough of an advantage to secure the victory in front of their home crowd . The Jazz were led by the duo of Derrick Favors and James Harden . Favors went 2 - for - 6 from the field and 0 - for - 1 from the three - point line to score a game - high of 15 points , while also adding four rebounds and four assists

Quantitative Evaluation / RotoWire

Model	Relations		Content		Ordering	
	P%	#	P%	R%	DLD%	BLEU
Template	99.35	49.7	45.17	24.85	12.2	6.87
Seq2Seq+Copy	71.07	12.61	21.90	27.27	8.70	14.46

What is Missing for Generation?

Building

Module	Content task	Structure task
Document planning	Content determination	Document structuring
Microplanning	Lexicalisation; Referring expression Generation	Aggregation
Realisation	Linguistic realisation	Structure realisation

Figure 3.1 Modules and tasks.

- Argument 1: Need bigger models. At greater than X parameters these properties will start showing up. (May likely be true)
- Argument 2: Just like copy we need richer models of the underlying phenomenon.

- 1 Preface: Seq2Seq+
- 2 Introduction: Data-Driven Text Generation
- 3 Latent-Variable Generation
- 4 Work 1: Learning Neural Templates
- 5 Work 2: Learning Alignments

Research Direction: Deep Latent-Variable Models for NLP

Goal: Expose specific choices as explicit *discrete* latent variables.

$$p(y, z; \theta).$$

- y is our observed data
- z is a collection of problem-specific latent variables
- θ are the deterministic, neural network parameters.

Research Direction: Deep Latent-Variable Models for NLP

Goal: Expose specific choices as explicit *discrete* latent variables.

$$p(y, z; \theta).$$

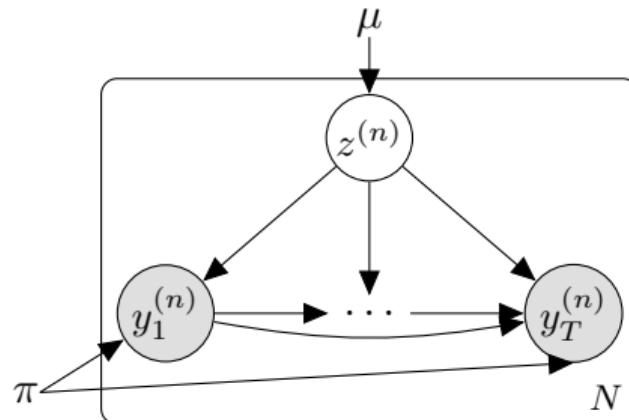
- y is our observed data
- z is a collection of problem-specific latent variables
- θ are the deterministic, neural network parameters.

Example Model: Mixture of RNNs

Generative process:

- ① Draw cluster $z \in \{1, \dots, K\}$ from a Categorical.
- ② Draw words $y_{1:T}$ from RNNLM with parameters π_z .

$$p(y, z; \theta) = \mu_z \times \text{RNNLM}(y_{1:T}; \pi_z)$$



Main Requirement: Posterior Inference

For models $p(y, z; \theta)$, we'll be interested in the *posterior* over latent variables z :

$$p(z | y; \theta) = \frac{p(y, z; \theta)}{p(y; \theta)} = \frac{p(y | z; \theta)p(z; \theta)}{\sum_{z'} p(y | z'; \theta)p(z'; \theta)}.$$

Why?

- Required for training
- Latent z gives separation of data.

How?

- Sum out over all discrete choices (e.g. run K RNNs).
- Variational inference based methods.

Main Requirement: Posterior Inference

For models $p(y, z; \theta)$, we'll be interested in the *posterior* over latent variables z :

$$p(z | y; \theta) = \frac{p(y, z; \theta)}{p(y; \theta)} = \frac{p(y | z; \theta)p(z; \theta)}{\sum_{z'} p(y | z'; \theta)p(z'; \theta)}.$$

Why?

- Required for training
- Latent z gives separation of data.

How?

- Sum out over all discrete choices (e.g. run K RNNs).
- Variational inference based methods.

In Applications: Copy-Attention (Gu et al, 2016) (Gulcehre et al, 2016)

Let z be a binary latent variable.

- If $z = 1$, let the model generate a new word.
- If $z = 0$, let the model copy a word from the source.

Inference:

Pointer-generator model + coverage summary

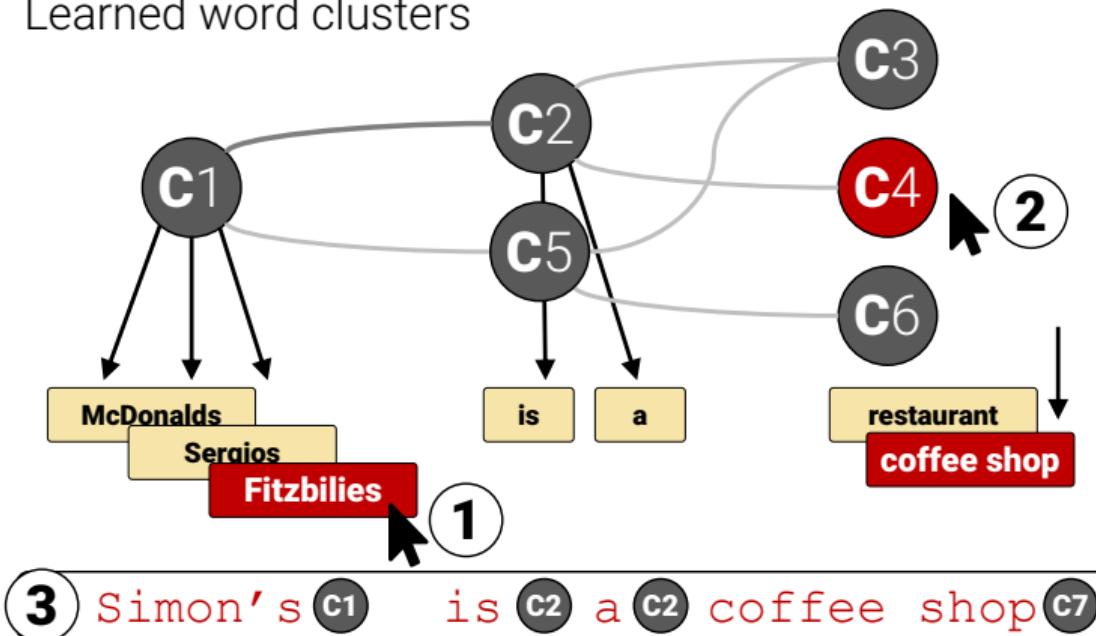
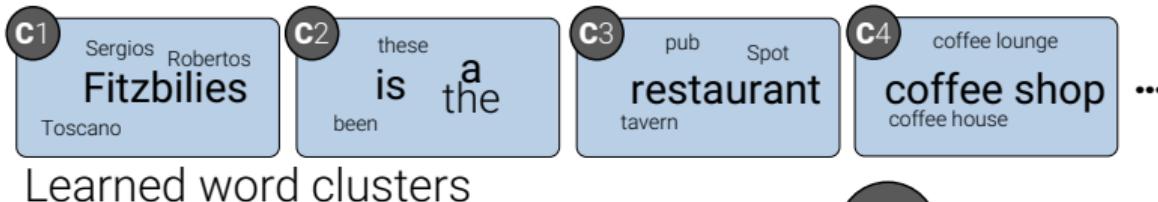
francis saili has signed a two-year deal to join munster later this year .
the 24-year-old was part of the new zealand under-20 side that won the junior world championship in italy in 2011 .
saili 's signature is something of a coup for munster and head coach anthony foley .

(See et al, 2017)

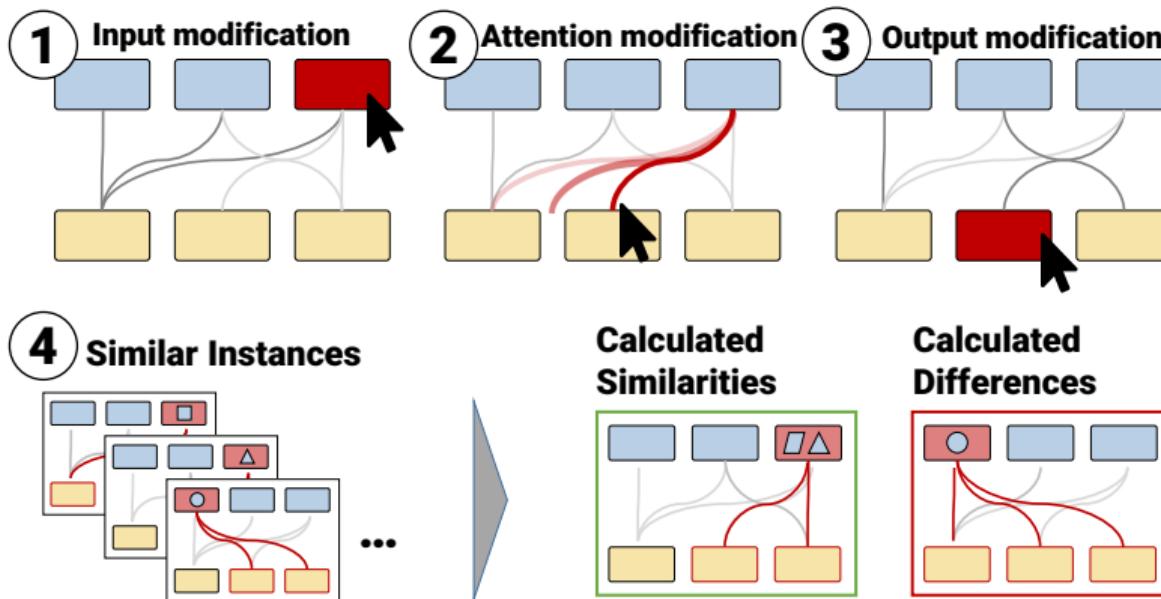
Latent Variable Models for Generation

- Can we develop other discrete latent-variable models for generation?
- Perhaps each important aspect of generation can be built-in directly.
- Goals:
 - Model Control
 - Model Debugging
 - Model Uncertainty

Approach 1: Learning Neural Templates



Approach 2: Latent Alignment and Variational Attention



- 1 Preface: Seq2Seq+
- 2 Introduction: Data-Driven Text Generation
- 3 Latent-Variable Generation
- 4 Work 1: Learning Neural Templates
- 5 Work 2: Learning Alignments

Text Generation: Talk about Structured Data (Generation)

TEAM	W	L	PTS	...
Heat	11	12	103	...
Hawks	7	15	95	...



Text Generation: Talk about Structured Data (Generation)

TEAM	W	L	PTS	...
Heat	11	12	103	...
Hawks	7	15	95	...



The Atlanta Hawks
defeated the Mi-
ami Heat, 103 - 95,
at Philips Arena on
Wednesday. Atlanta
...

Template-style Text Generation

A classical NLG template:

```
<restaurant_name> is a  
<food_type> <restaurant_type>  
with a <num_stars> star rating. It is  
located in <neighborhood>, and its  
price range is <price_range>.
```

Why Templates?

Template-based generation addresses certain deficiencies in encoder/decoder style generation.

- More interpretable
- More controllable

Where We're Going

An end-to-end, encoder/decoder style model that allows for:

- Induction of discrete, template-like objects from text
- Interpretable and controllable generation with these induced templates
- Good, but not quite SOTA performance on automatic metrics

Data-to-Text Generation

[c.f., Lebret et al., 2016]

Frederick Parker-Rhodes

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



“Frederick Parker-Rhodes
(21 November 1914 – 2
March 1987) was an
English linguist, plant
pathologist, computer
scientist, mathematician,
mystic, and mycologist.”

Data-to-Text Generation

[c.f., Novikova et al., 2017]

Name	The Eagle
Eat Type	coffee shop
Food	French
Price Range	moderate
Customer Rating	3/5
Area	riverside
Kids Friendly	yes
Near	Burger King



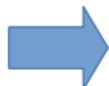
“The three star coffee shop, The Eagle, gives families a mid-priced dining experience featuring a variety of wines and cheeses. Find The Eagle near Burger King.”

Argument for Templates #1: Interpretability

Frederick Parker-Rhodes

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

Frederick Parker-Rhodes (21 November 1914 – 2 March 1987) was an English mycology and plant pathology, mathematics at the University of UK.”

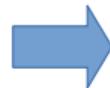


Argument for Templates #1: Interpretability

Frederick Parker-Rhodes

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

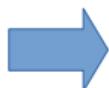
Frederick Parker-Rhodes (21 November 1914 – 2 March 1987) was an English mycology and plant pathology, mathematics at the University of UK.”



<name> (born <born>) was a
<nationality> <occupation>, who
lived in the <residence>. He was
known for contributions to <known_for>.

Argument for Templates #2: Controllability

Name	The Eagle
Eat Type	coffee shop
Food	French
Price Range	moderate
Customer Rating	3/5
Area	riverside
Kids Friendly	yes
Near	Burger King



<name> is a kid-friendly <eat_type> serving <food> cuisine in the <area> area.

The <customer_rating> star rated <name> serves <food> food at a <price_range> price.

Near <near> is a <food> <eat_type> with a <customer_rating> star rating. It is family friendly, and its price range is <price_range>.

Goal: Learned Template-style Generation

- **Idea:** use a Hidden Semi-Markov Model (HSMM) decoder
 - Preserve most of the encoder/decoder setup
 - Learn template-like representations jointly with learning to generate

Hidden Semi-Markov Models

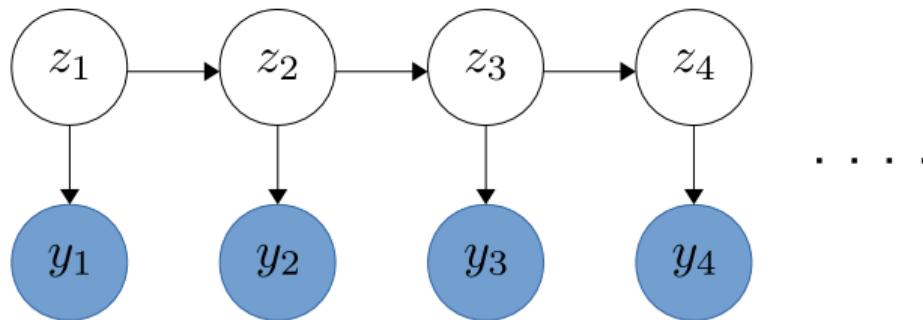
[Gales and Young, 1993; Ostendorf et al., 1996]

- Give a joint distribution over observations $y_{1:T}$ and discrete latents $z_{1:S}$
 - Like HMMs, but observations can last multiple time-steps:

Hidden Semi-Markov Models

[Gales and Young, 1993; Ostendorf et al., 1996]

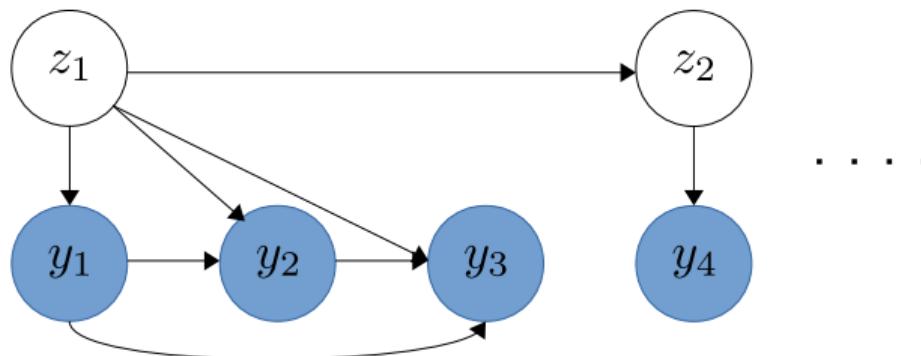
- Give a joint distribution over observations $y_{1:T}$ and discrete latents $z_{1:S}$
 - Like HMMs, but observations can last multiple time-steps:



Hidden Semi-Markov Models

[Gales and Young, 1993; Ostendorf et al., 1996]

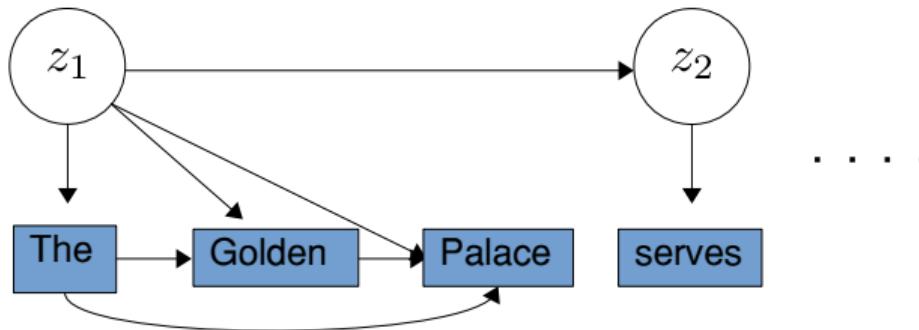
- Give a joint distribution over observations $y_{1:T}$ and discrete latents $z_{1:S}$
 - Like HMMs, but observations can last multiple time-steps:



Hidden Semi-Markov Models

[Gales and Young, 1993; Ostendorf et al., 1996]

- Give a joint distribution over observations $y_{1:T}$ and discrete latents $z_{1:S}$
 - Like HMMs, but observations can last multiple time-steps:

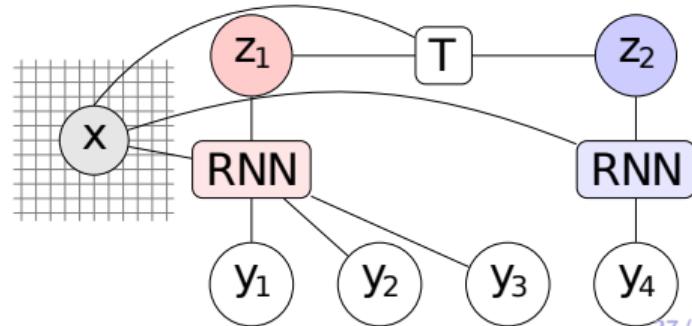


Upshot: HSMMs give us typed segmentations.

A Conditional (Neural) HSMM

$$p(y, z | x) = \prod_{s=1}^S \underbrace{p(z_s | z_{s-1}, x)}_{\text{transition prob}} \underbrace{p(l_s | z_s)}_{\text{length prob}} \underbrace{p(y_{t_0(s):t_1(s)} | z_s, l_s, x)}_{\text{segment prob}}$$

- We parameterize probabilities with neural components:
 - Segment probabilities are given by an RNN + attention + copy attention



Learning

- We're given a dataset of x, y pairs
- Segmentations z are unobserved at training time
- **Maximize** $\ln p(y_{1:T} | x) = \ln \sum_z p(y_{1:T}, z | x)$
 - Can use a dynamic program analogous to the forward or backward algorithm used in learning HMMs [c.f., Murphy 2002]
 - Can simply backprop through the dynamic program
 - Easy with pytorch!

Generation

- Given an input x , we could generate by approximating $\arg \max_{y,z} p(y, z | x)$
- Instead, we'll first extract "templates":
 - 1) Viterbi-segment the training data

Generation

- Given an input x , we could generate by approximating $\arg \max_{y,z} p(y, z | x)$
- Instead, we'll first extract "templates":
 - 1) Viterbi-segment the training data

```
[The Golden Palace]55 [is a]59 [coffee shop]12 [providing]3 [Indian]50
[food]1 [in the]17 [ $\text{£}20\text{-}25$ ]125 [price range]16 [.]2 [It is]8 [located
in the]25 [riverside]40 [.]53 [Its customer rating is]19 [high]23 [.]2
```

Generation

- Given an input x , we could generate by approximating $\arg \max_{y,z} p(y, z | x)$
- Instead, we'll first extract "templates":
 - 1) Viterbi-segment the training data

```
[The Golden Palace]55 [is a]59 [coffee shop]12 [providing]3 [Indian]50
[food]1 [in the]17 [ $\text{£}20\text{-}25$ ]125 [price range]16 [.]2 [It is]8 [located
in the]25 [riverside]40 [.]53 [Its customer rating is]19 [high]23 [.]2
```

Note: each segment gets a latent-state label.

- We'll call a sequence of labels $z^{(i)}$ a "template."
- E.g., $z^{(i)} = 55, 59, 12, 3, \dots$

What's Good about these “Templates”?

1) Dim. reduction: latent states correspond to functional categories.

1. The Eagle provides Indian food in the high price range. It is located near riverside city centre. Its customer rating is 1 out of 5. It has a average rating. They are located in the near It's located near Cafe Sicilia. The price range is high.
2. Located near The Portland Arms is an Italian restaurant called The Waterman. Located in the riverside city centre. It is a family friendly fast food place called Cocum. There is a French restaurant named Loch Fyne.
3. A Italian restaurant is The Waterman. An fast food pub coffee shop is called Cocum. French named Loch Fyne.
4. Located near The Portland Arms. Located in the riverside city centre. The Eagle is a cheap Italian restaurant. The Golden Curry is a family-friendly fast food pub coffee shop. Zizzi is an family friendly French restaurant.
5. A Italian restaurant is The Waterman. An fast food pub coffee shop is called Cocum. French named Loch Fyne. Located near riverside city centre. Located in the located near Cafe Sicilia.

What's Good about these “Templates”?

1) Dim. reduction: latent states correspond to functional categories.

1. aftab ahmed | born 1951 | is an american actor
anderson da silva | (born on 1970 | was an american actress | .
david jones | ; born 1974 | is an english cricketer | .
...
2. aftab ahmed | was a world war i member of the austrian house of representatives
anderson da silva | is a former liberal party member of the pennsylvania legislature
david jones | is a baseball recipient of the montana senate | .
...
3. adjutant | aftab ahmed | was a world war i member of the knesset
lieutenant | anderson da silva | is a former liberal party member of the scottish parliament | .
captain | david jones | is a baseball recipient of the fc lokomotiv liski | .
...
4. william | " billy " watson | 1913 | 1917 | was an american football player
john william | smith | (c. 1900 | in surrey, england | was an australian rules footballer
james " | jim " edward | 1913 | - | is an american defenceman | .
...
5. who plays for | collingwood | in the victorial football league | vfl
who currently plays for | st kilda | of the national football league | afl
who played with | carlton | and the australian football league | nfl | .
...
6. aftab ahmed | is a member of the knesset
anderson da silva | is a former party member of the scottish parliament | .
david jones | is a female recipient of the fc lokomotiv liski | .
...

What's Good about these “Templates”?

2) We can use them to control generation:

- Select a template $z^{(i)} = z_1^{(i)}, \dots, z_S^{(i)}$
- Generate by computing $\arg \max_y p(y, z = z^{(i)} | x)$
- Gives a different generation for each $z^{(i)}$
 - (Examples in a few slides...)

Generation Recap

- 1) Viterbi-segment the training data
- 2) Collect frequent “templates” $z^{(i)} = z_1^{(i)}, \dots, z_S^{(i)}$
- 3) Given a new input x , generate by finding $\arg \max_y p(y, z^{(i)} | x)$ for a chosen template $z^{(i)}$

Methods

- 1) Condition RNNs on latent state by concatenating state-embedding to RNN input
- 2) Helpful to train with hard constraints: disallow splitting up segments appearing in tables
- 3) Segment RNNs can condition on all preceding *tokens*

E2E Validation Results

(Val)	BLEU	NIST	ROUGE	CIDEr	METEOR
D&J (2017)	69.25	8.48	72.57	2.40	47.03
Substitution BL					
Neural Template					

- D&J (2017) is an enc/dec + reranker system used in the E2E Challenge
- Substitution BL finds maximally similar training table and performs substitution in corresponding description
- K=60; 1x300 LSTM as segment models
- Used 100 most common $z^{(i)}$ and selected highest overall scorer

E2E Validation Results

(Val)	BLEU	NIST	ROUGE	CIDEr	METEOR
D&J (2017)	69.25	8.48	72.57	2.40	47.03
Substitution BL	43.71	6.72	55.35	1.41	37.87
Neural Template					

- D&J (2017) is an enc/dec + reranker system used in the E2E Challenge
- Substitution BL finds maximally similar training table and performs substitution in corresponding description
- K=60; 1x300 LSTM as segment models
- Used 100 most common $z^{(i)}$ and selected highest overall scorer

E2E Validation Results

(Val)	BLEU	NIST	ROUGE	CIDEr	METEOR
D&J (2017)	69.25	8.48	72.57	2.40	47.03
Substitution BL	43.71	6.72	55.35	1.41	37.87
Neural Template	67.07	7.98	69.50	2.29	43.07

- D&J (2017) is an enc/dec + reranker system used in the E2E Challenge
- Substitution BL finds maximally similar training table and performs substitution in corresponding description
- K=60; 1x300 LSTM as segment models
- Used 100 most common $z^{(i)}$ and selected highest overall scorer

E2E Test Results

(Val)	BLEU	NIST	ROUGE	CIDEr	METEOR
D&J (2017)	65.93	8.59	68.50	2.23	44.83
Substitution BL	43.78	6.88	54.64	1.39	37.35
Neural Template	59.80	7.56	65.01	1.95	38.75

- D&J (2017) is an enc/dec + reranker system used in the E2E Challenge
- Substitution BL finds maximally similar training table and performs substitution in corresponding description
- K=60; 1x300 LSTM as segment models
- Used 100 most common $z^{(i)}$ and selected highest overall scorer

WikiBio Results

	BLEU	NIST	ROUGE-4
Template KN	19.8	5.19	10.7
NNLM (field)	33.4	7.52	23.9
NNLM (field & word)	34.7	7.98	25.8
Neural Template	34.8	7.59	38.6

- Encoder/decoder and template-style baselines from Lebret et al. (2016)
- K=45; 1x300 LSTMs as segment/history models
- Used 100 most common $z^{(i)}$ and selected highest overall scorer

WikiBio Results

	BLEU	NIST	ROUGE-4
Template KN	19.8	5.19	10.7
NNLM (field)	33.4	7.52	23.9
NNLM (field & word)	34.7	7.98	25.8
Neural Template	34.8	7.59	38.6
Liu et al. (2018)	43.7	-	40.3

- Encoder/decoder and template-style baselines from Lebret et al. (2016)
- K=45; 1x300 LSTMs as segment/history models
- Used 100 most common $z^{(i)}$ and selected highest overall scorer

Controllability Example (E2E)

$z = 55, 59, 43, 11, 25, 50, 53$

Travellers Rest Beefeater₅₅ is a₅₉ 3 star₄₃ restaurant₁₁ located near₂₅ Raja Indian Cuisine_{40 · 53}

Name	Travellers Rest Beefeater
Customer Rating	3 out of 5
Area	riverside
Near	Raja Indian Cuisine

Controllability Example (E2E)

$z = 31, 29, 44, 55, 3, 50,$
 $1, 2$

Name	Travellers Rest Beefeater
Customer Rating	3 out of 5
Area	riverside
Near	Raja Indian Cuisine

Travellers Rest Beefeater₅₅ is a₅₉ 3 star₄₃ restaurant₁₁ located near₂₅ Raja Indian Cuisine_{40 · 53}

Near₃₁ riverside_{29 · 44} Travellers Rest Beefeater₅₅ serves₃ 3 star₅₀ food_{1 · 2}

Controllability Example (E2E)

$z = 55, 59, 12, 3, 50, 1, 17,$
 $26, 16, 2, 8, 25, 40, 53$

Name	Travellers Rest Beefeater
Customer Rating	3 out of 5
Area	riverside
Near	Raja Indian Cuisine

Travellers Rest Beefeater₅₅ is a₅₉ 3 star₄₃ restaurant₁₁ located near₂₅ Raja Indian Cuisine_{40 · 53}

Near₃₁ riverside_{29 · 44} Travellers Rest Beefeater₅₅ serves₃ 3 star₅₀ food_{1 · 2}

Travellers Rest Beefeater₅₅ is a₅₉ restaurant₁₂ providing₃ riverside₅₀ food₁ and has a₁₇ 3 out of 5₂₆ customer rating_{16 · 2}. It is₈ near₂₅ Raja Indian Cuisine_{40 · 53}

Controllability Example (E2E)

Name	Travellers Rest Beefeater
Customer Rating	3 out of 5
Area	riverside
Near	Raja Indian Cuisine

Travellers Rest Beefeater₅₅ is a₅₉ 3 star₄₃ restaurant₁₁ located near₂₅ Raja Indian Cuisine_{40 · 53}

Near₃₁ riverside_{29 · 44} Travellers Rest Beefeater₅₅ serves₃ 3 star₅₀ food_{1 · 2}

Travellers Rest Beefeater₅₅ is a₅₉ restaurant₁₂ providing₃ riverside₅₀ food₁ and has a₁₇ 3 out of 5₂₆ customer rating_{16 · 2}. It is₈ near₂₅ Raja Indian Cuisine_{40 · 53}

Travellers Rest Beefeater₅₅ is a₅₉ place to eat₁₂ located near₂₅ Raja Indian Cuisine_{40 · 53}

Travellers Rest Beefeater₅₅ is a₅₉ 3 out of 5₅ rated₃₂ riverside₄₃ restaurant₁₁ near₂₅ Raja Indian Cuisine_{40 · 53}

Interpretability Example (WikiBio)

Jimmy Deacon

Personal information		
Full name	James Deacon	
Date of birth	23 January 1906	
Place of birth	Glasgow, Scotland	
Date of death	1976 (aged 69–70)	
Height	5 ft 7 in (1.70 m)	
Playing position	Forward	
Senior career*		
Years	Team	Apps (Gls)
	Darlington	2 (-)
1929–1934	Wolverhampton Wanderers	149 (52)
1934–1939	Southend United	100 (3)
1939–1940	Hartlepool	- (-)

* Senior club appearances and goals counted for the domestic league only

Yang Sung-chul

Born	20 November 1939 (age 78) Gokseong County, Jeollanam-do
Citizenship	South Korea
Alma mater	Seoul National University University of Hawaii at Manoa University of Kentucky
Occupation	Political scientist
Employer	Graduate School of International Studies, Korea University
Known for	Member of the National Assembly Ambassador to the United States
Political party	National Congress for New Politics
Children	Two

james deacon₄₂ (₃₀ born₄₄ 23 january 1906₁₁)₂₂
was a₁₄ scottish₈ football₁₉ forward₂₄ · 43

yang sung-chul₄₂ (₃₀ november 20, 1939₁₁ in₂₁ gokseong county, jeollanam-do₃₉)₂₂ is a₁₄ south korean₈ political scientist₂₄ · 43

- 1 Preface: Seq2Seq+
- 2 Introduction: Data-Driven Text Generation
- 3 Latent-Variable Generation
- 4 Work 1: Learning Neural Templates
- 5 Work 2: Learning Alignments

Text Generation: Talk about Text (Translation / Summarization)

mexico city , mexico -lrb- cnn -rrb- – heavy rains and flooding have forced hundreds of thousands of people from homes in southern mexico 's state of tabasco over the past four days , with nearly as many trapped by the rising waters , state officials said thursday . officials say about 300,000 people are still trapped by the worst flooding in the region for 50 years . the grijalva river pushed over its banks through the state capital of villahermosa on thursday , forcing government workers to evacuate and leaving up to 80 percent of the city flooded , gov. andres granier 's office told cnn . about 700,000 people have seen their homes flooded , with about 300,000 of those still trapped there , granier 's office reported . one death had been blamed on the floods , which followed weeks of heavy rain in the largely swampy state . tabasco borders guatemala to the south and the gulf of mexico to the north . . .



Text Generation: Talk about Text (Translation / Summarization)

mexico city , mexico -lrb- cnn -rrb- – heavy rains and flooding have forced hundreds of thousands of people from homes in southern mexico 's state of tabasco over the past four days , with nearly as many trapped by the rising waters , state officials said thursday . officials say about 300,000 people are still trapped by the worst flooding in the region for 50 years . the grijalva river pushed over its banks through the state capital of villahermosa on thursday , forcing government workers to evacuate and leaving up to 80 percent of the city flooded , gov. andres granier 's office told cnn . about 700,000 people have seen their homes flooded , with about 300,000 of those still trapped there , granier 's office reported . one death had been blamed on the floods , which followed weeks of heavy rain in the largely swampy state . tabasco borders guatemala to the south and the gulf of mexico to the north . . .



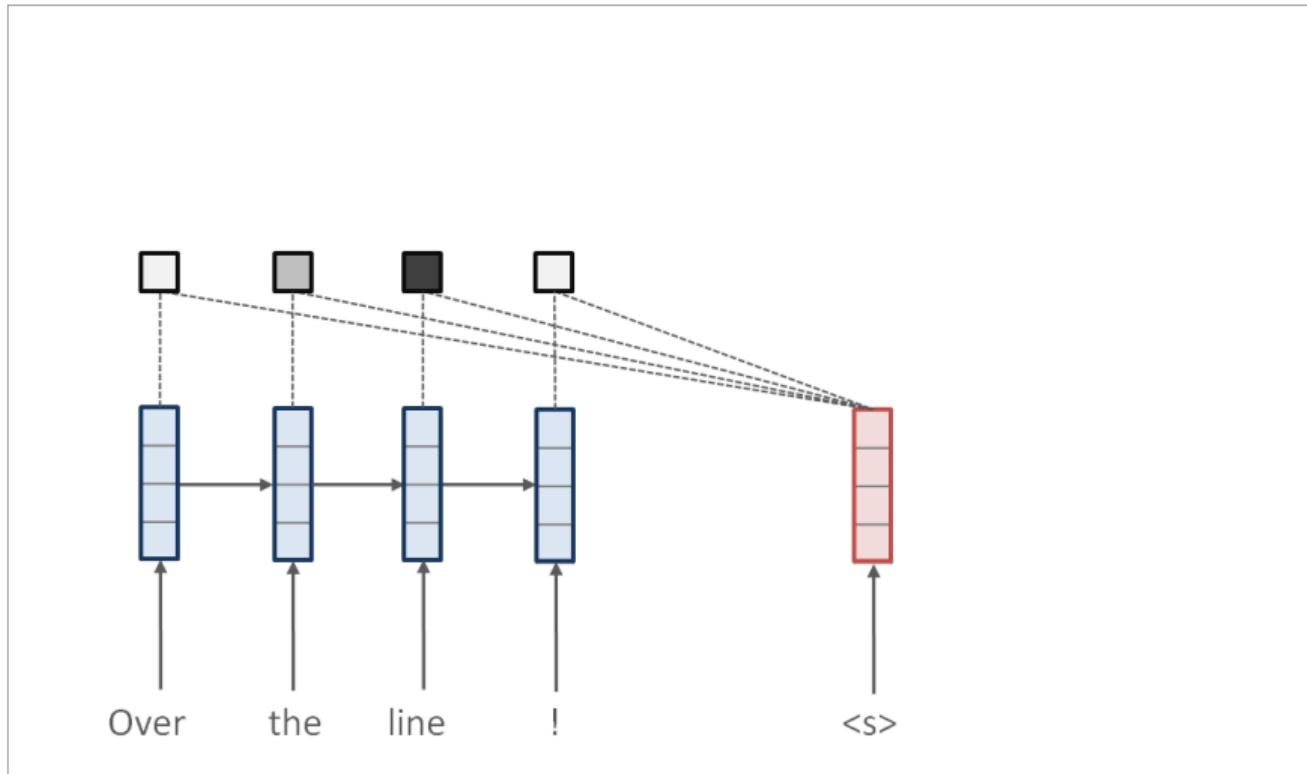
tabasco and chia-
pas states hardest
hit. authorities say
700,000 affected . . .

Motivation:
Six Challenges for NMT (Koehn and Knowles 2017)

- 2: Requires large sample complexity
- 5: The alignments learned by soft attention may not be interpreted as word alignments

These issues are even greater for text generation.

Seq2Seq+



Neural Attention versus Latent Alignment

- Attention is motivated as an approximation to alignment:
 - makes the model's behavior interpretable
 - can be used for other prediction tasks
- We contrast this with latent alignment which acts as a random variable.

Key Distinction

- Both can compute $p(x|z)$ but attention cannot compute posterior $p(z|x)$.

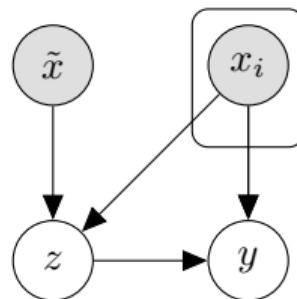
Latent Alignment: Motivation

If attention works so well, why study alignment?

- A latent variable approach facilitates **composability** in a principled probabilistic manner.
(Cohn et al, 2016)
- **Posterior inference** provides better post-hoc interpretability and analysis
- Modeling **uncertainties** might lead to better performance

Notation: Alignment Model

- $x = x_1, \dots, x_T$: the observed set (the encoded source words and previous target words)
- \tilde{x} : the query (the decoder hidden state at a single timestep)
- y : the output (the current target word)
- z : the latent alignment, random variable indicating which member of x generates y



Problem Setup

- Let \mathcal{D} be the prior distribution of z and $f(x, z; \theta)$ the likelihood of x given z
- Generative Process

$$z \sim \mathcal{D}(a(x, \tilde{x}; \theta)) \quad y \sim f(x, z; \theta)$$

- Training Objective (maximizing marginal log-likelihood)

$$\max_{\theta} \log p(y = \hat{y} | x, \tilde{x}) = \max_{\theta} \log \mathbb{E}_z[f(x, z; \theta) \hat{y}]$$

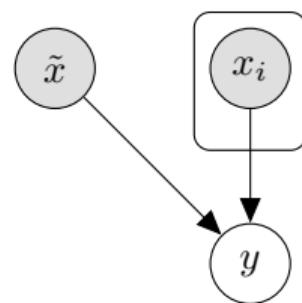
- Direct optimization is computationally expensive
 - Discrete $z \sim \mathcal{D}$: $O(T)$ additional runtime
 - Continuous $z \sim \mathcal{D}$: intractable

Workaround 1: Soft Attention

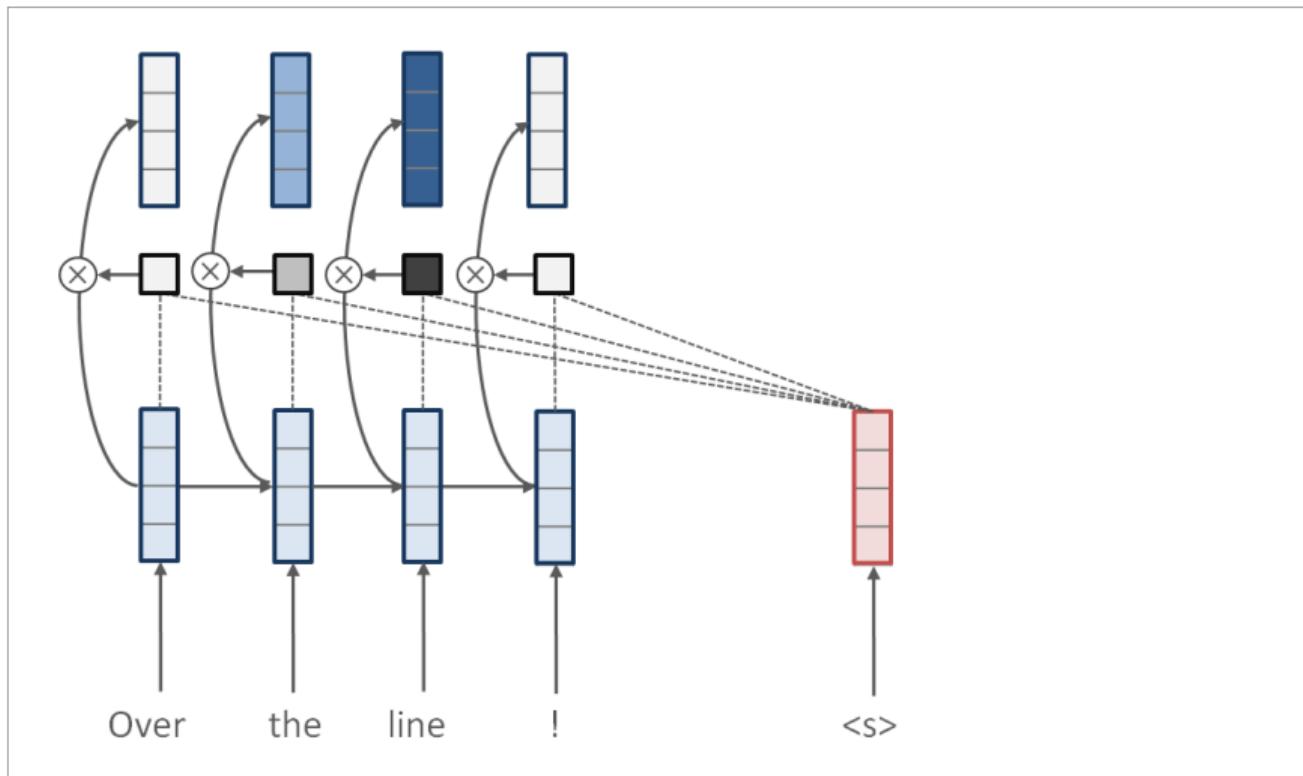
- Replace the joint distribution with a nested expectation [Bahdanau et al 2014]

$$\log \mathbb{E}_z[f(x, z)_{\hat{y}}] \approx \log f(x, \mathbb{E}_z[z])_{\hat{y}}$$

- The corresponding graphical model is



Seq2Seq+



Workaround 2: Hard Attention

- Keep the latent variable model formulation and maximize a lower bound on the marginal likelihood
- [Xu et al 2015]: Directly apply Jensen's inequality and optimize with REINFORCE by sampling from the prior

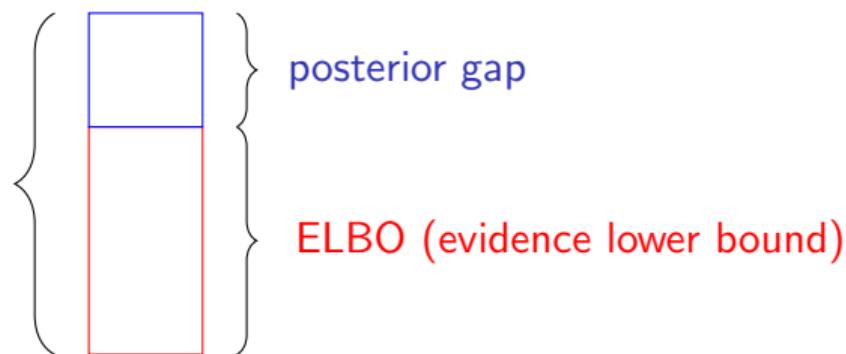
$$\log \mathbb{E}_z[f(x, z)_{\hat{y}}] \geq \mathbb{E}_z \log[f(x, z)_{\hat{y}}] \approx \log f(x, \tilde{z})_{\hat{y}}$$

- The use of the prior in the expectation may result in a poor bound
- Cannot directly use for posterior estimation.

Marginal Likelihood: Variational Decomposition

For any¹ distribution $q(z)$ over z ,

$$L(\theta) = \mathbb{E}_q \left[\log p(y | x, z) \right] - \text{KL}[q(z) \| p(z | x, \tilde{x})] \\ + \text{KL}[q(z) \| p(z | y, x, \tilde{x})]$$



Since KL is always non-negative, $L(\theta) \geq \text{ELBO}(\theta, \lambda)$.

¹Technical condition: $\text{supp}(q(z)) \subset \text{supp}(p(z | x; \theta))$

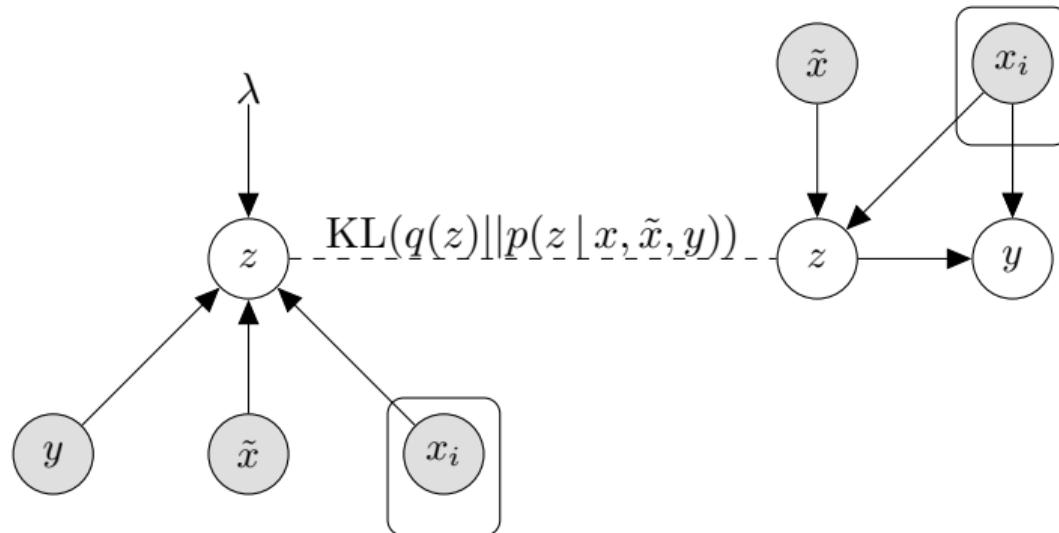
Proposal: Variational Attention

- Learn model and q to maximize the following lower bound

$$\begin{aligned} & \log \mathbb{E}_{z \sim p(z|x, \tilde{x})} [p(y|x, z)] \\ & \geq \mathbb{E}_{z \sim q(z)} [\log p(y|x, z)] - \text{KL}[q(z) \| p(z|x, \tilde{x})] \end{aligned}$$

- We choose a pair \mathcal{D} and $q(z)$ that affords analytic KL
- At test time, marginalize over z during decoding

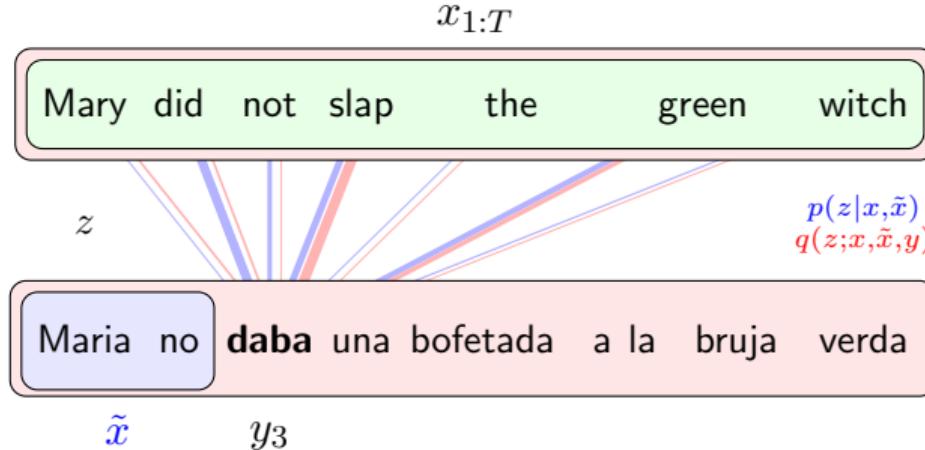
Example Form of q : Amortized Parameterization



λ parameterizes a global network (encoder) that is run over x, y, \tilde{x} to produce the local variational distribution, e.g.

$$q(z; \lambda) = \text{enc}(x, \tilde{x}, y; \lambda)$$

Method



A sketch of variational attention applied to translation.

- The blue prior p is restricted to past information,
- The red variational posterior q may take into account future observations.

Proposal: Categorical and Relaxed

- Categorical (Single Source Alignment Word)

- $z \sim \mathcal{D}$ and $q(z)$: Categorical
- Estimate gradients with REINFORCE

$$\mathbb{E}_{z \sim q(z)} [\nabla_{\theta} \log f(x, z) + \log f(x, z) \nabla_{\phi} \log q(z)]$$

- Relaxed (Mixture Source Alignment)

- $z \sim \mathcal{D}$ and $q(z)$: Dirichlet
- Use the reparameterization trick [Kingma et al 2013]
 - Sample u from a simple distribution \mathcal{U}
 - Apply transformation $g_{\phi}(\cdot)$ to obtain $z = g_{\phi}(u)$
- The gradient estimator takes the form

$$\mathbb{E}_{u \sim \mathcal{U}} [\nabla_{\theta, \phi} \log f(x, g_{\phi}(u))]$$

Proposal: Categorical and Relaxed

- Categorical (Single Source Alignment Word)

- $z \sim \mathcal{D}$ and $q(z)$: Categorical
 - Estimate gradients with REINFORCE

$$\mathbb{E}_{z \sim q(z)} [\nabla_{\theta} \log f(x, z) + \log f(x, z) \nabla_{\phi} \log q(z)]$$

- Relaxed (Mixture Source Alignment)

- $z \sim \mathcal{D}$ and $q(z)$: Dirichlet
 - Use the reparameterization trick [Kingma et al 2013]
 - Sample u from a simple distribution \mathcal{U}
 - Apply transformation $g_{\phi}(\cdot)$ to obtain $z = g_{\phi}(u)$
 - The gradient estimator takes the form

$$\mathbb{E}_{u \sim \mathcal{U}} [\nabla_{\theta, \phi} \log f(x, g_{\phi}(u))]$$

Important Extension: Variance Reduction for Categorical

- REINFORCE gradient estimator suffers from high variance
- Introduce control variate or baseline $B = \log f(x, \mathbb{E}_{z' \sim q(z)}[z'])$ from soft attention

$$\mathbb{E}_{z \sim q(z)}[\nabla_{\theta} \log f(x, z) + (\log f(x, z) - B) \nabla_{\phi} \log q(z)]$$

- Requires a single additional evaluation of $f(x, \mathbb{E}_{z' \sim q(z)}[z'])$

Concurrent Experimental Work

Many different researchers have recently explored the benefits of marginalization. Very similar results.

- Surprisingly Easy Hard-Attention for Sequence to Sequence Learning
- Hard Non-Monotonic Attention for Character-Level Transduction
- Posterior Attention Models for Sequence to Sequence Learning

These results seem robust across these works.

Experiments

- Full experiments on IWSLT and WMT using LSTM based NMT system.
- Model: Two layer attention based LSTM.
- Variational Model: Bidirectional LSTM model.
- Preliminary experiments on low-resource MT setup.

Results (MT: IWSLT)

Model	Objective	Exp	PPL	BLEU
Soft Attn	$\log p(y \mathbb{E}[z])$	Softmax	7.17	32.77
Marg. Likelihood	$\log \mathbb{E}[p]$	Enum	6.34	33.29
Hard Attn	$\mathbb{E}_p[\log p]$	Enum	6.77	31.40
Hard Attn	$\mathbb{E}_p[\log p]$	Sample	6.78	30.42
Var Relaxed Attn	$\mathbb{E}_q[\log p] - \text{KL}$	Sample	7.58	30.05
Var Attn	$\mathbb{E}_q[\log p] - \text{KL}$	Enum	6.08	33.69
Var Attn	$\mathbb{E}_q[\log p] - \text{KL}$	Sample	6.17	33.30

Results (MT: WMT)

Model	Objective	Exp	PPL	BLEU
Soft Attn	$\log p(y \mid \mathbb{E}[z])$	Softmax	-	24.10
Var Attn	$\mathbb{E}_q[\log p] - \text{KL}$	Sample	-	24.98

Preliminary Results (Low Data Settings)

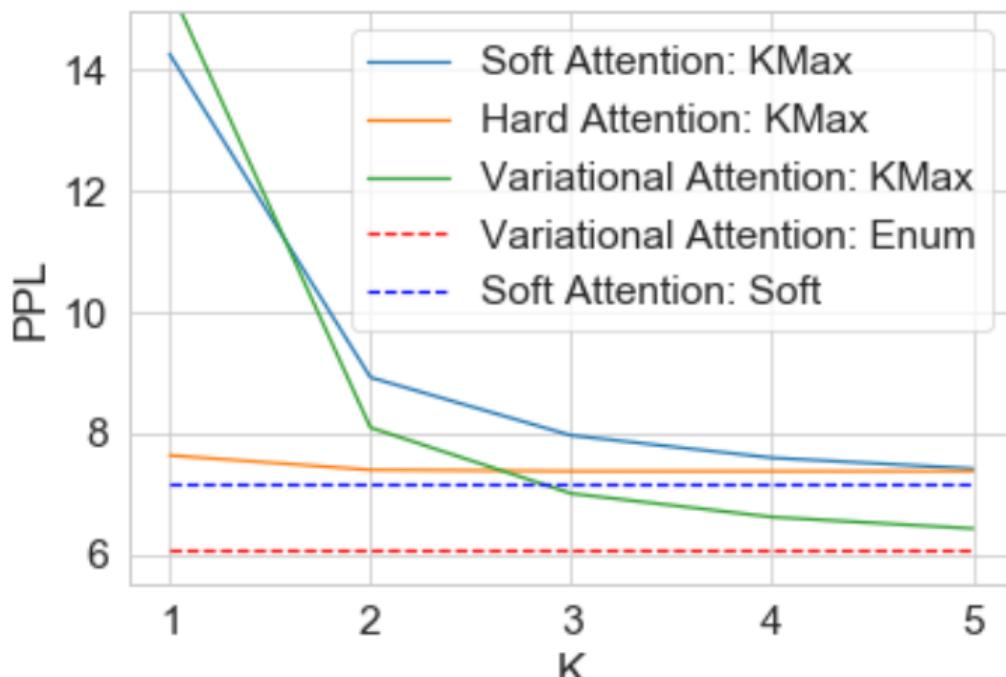
Training Size	10k	25k	50k	Full (160k)
Soft Attn	12.10	22.88	26.65	32.77
Marginal Likelihood	16.79	23.44	26.99	33.29
Var Attn Enum	14.90	23.50	27.26	33.69
Var Attn Sample	12.20	23.35	27.87	33.30

Results (VQA)

Model	Objective	Exp	NLL	Eval
Soft Attn	$\log p(y \mathbb{E}[z])$	Softmax	1.76	58.93
Marg. Likelihood	$\log \mathbb{E}[p]$	Enum	1.69	60.33
Hard Attn	$\mathbb{E}_p[\log p]$	Enum	1.78	57.60
Hard Attn	$\mathbb{E}_p[\log p]$	Sample	1.82	56.30
Var Attn	$\mathbb{E}_q[\log p] - \text{KL}$	Enum	1.68	58.44
Var Attn	$\mathbb{E}_q[\log p] - \text{KL}$	Sample	1.74	57.52

The Exp column indicates whether enumeration (Enum) or sampling (Sample) was used during training. We evaluate intrinsically on negative log-likelihood NLL (lower is better) and VQA evaluation metric (higher is better).

Inference



Discussion

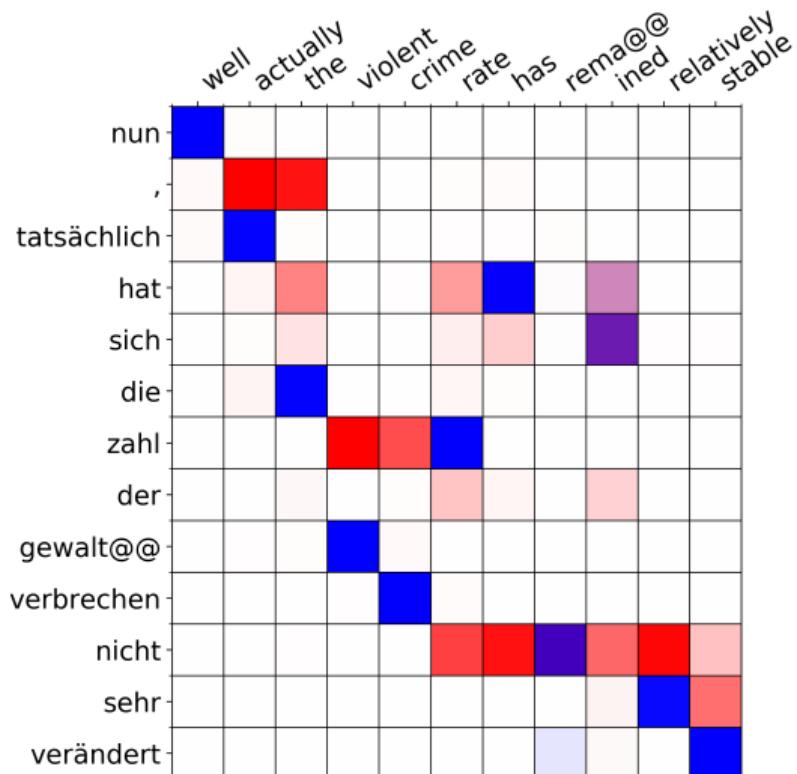
- Soft attention underperforms latent attention
- Variational attention achieves similar performance to maximizing the marginal likelihood exactly despite optimizing a lower bound
- Proposed soft attention as a computationally cheap baseline for reducing the variance of the REINFORCE gradient estimator

Discussion: Alternative Inference Methods

Inference Method	#Samples	PPL	BLEU
REINFORCE	1	6.17	33.30
RWS	5	6.41	32.96
Gumbel-Softmax	1	6.51	33.08

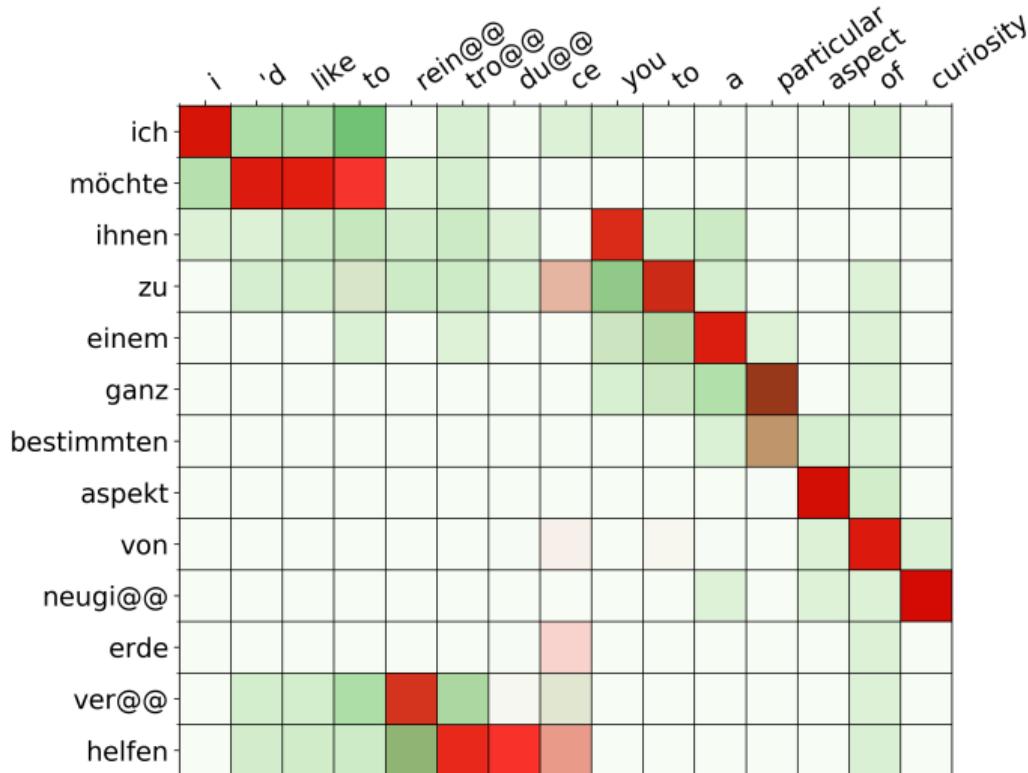
- Gumbel-Softmax is a viable alternative
- RWS incurs higher memory cost

Example Alignments



Red: prior; blue: posterior.

Example Alignments



Red: prior; green: soft attention.

Conclusion

- Discrete latent variables are a natural fit for NLP
- Methods can be integrated naturally into many problems.
- Tools for training these approaches are coming along fast.
- Also check out our tutorial: Deep Latent-Variable NLP bit.do/lvnlp