Controlling Text Generation

Alexander Rush

(with Yoon Kim, Sam Wiseman, Sebastian Gehrmann,

Yuntian Deng, Justin Chiu, and Demi Guo)

Harvard 2018

nlp.seas.harvard.edu

1. Introduction: Data-Driven Text Generation
2. Latent-Variable Generation
3. Work 1: Learning Neural Templates
4. Work 2: Learning Alignments
5. Intro

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 chiapas 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	...

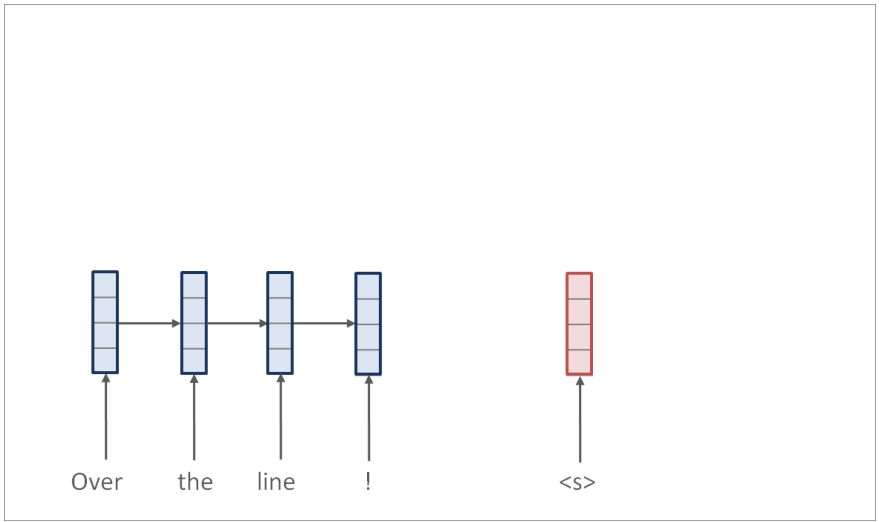
The Atlanta Hawks defeated the Miami Heat, 103 - 95, at Philips Arena on Wednesday. Atlanta ...

Text Generation:
Talk about the Environment (Multimodal)

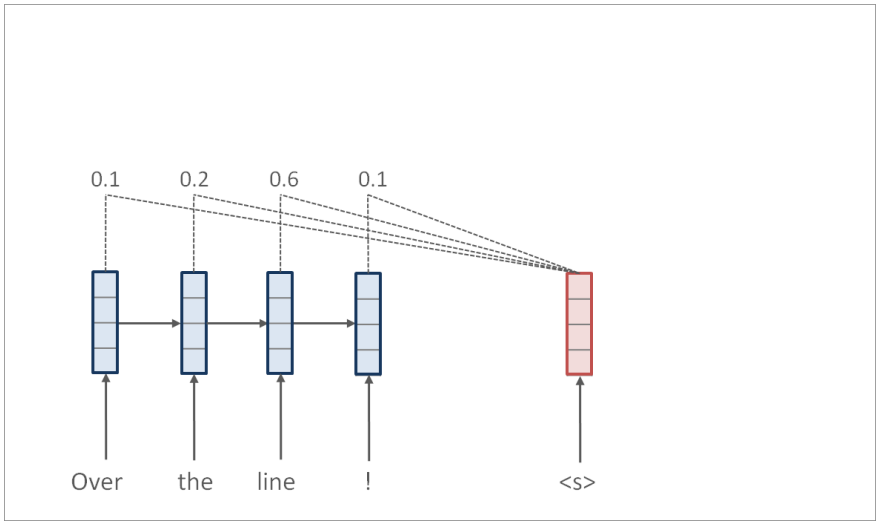
Text Generation:
Talk about the Environment (Multimodal)

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

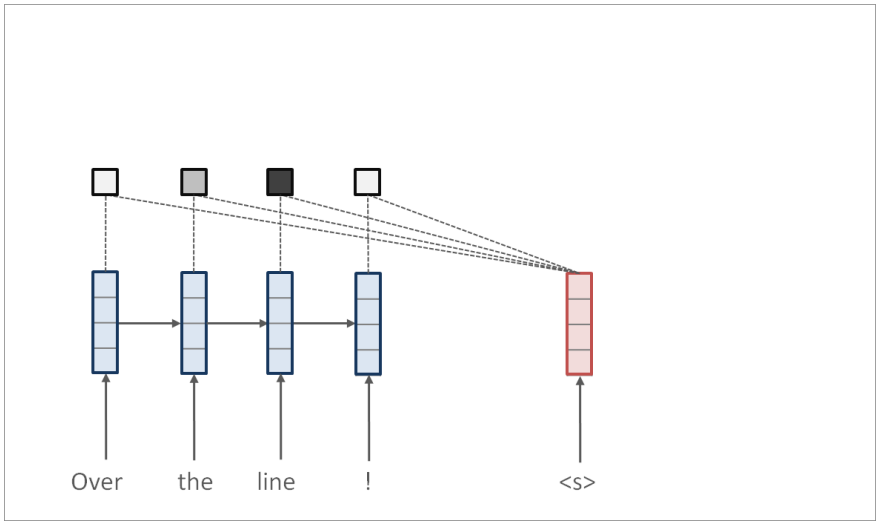
Seq2Seq+



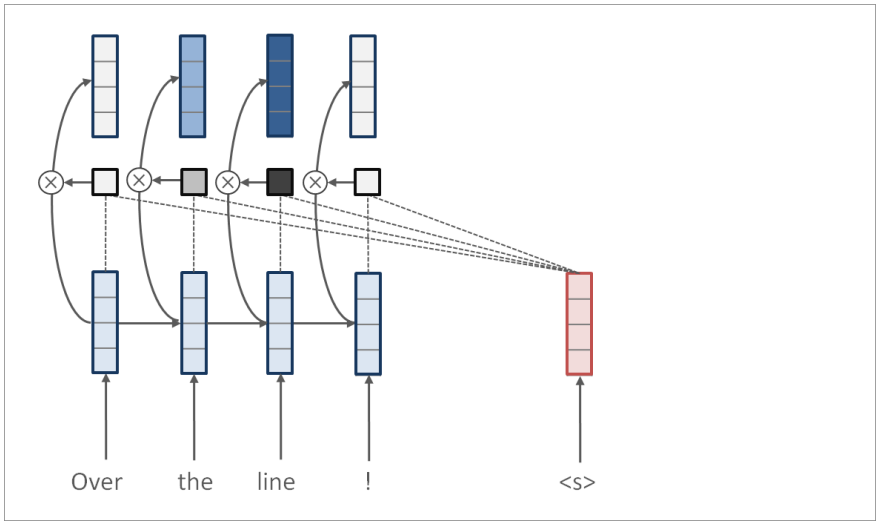
Seq2Seq+



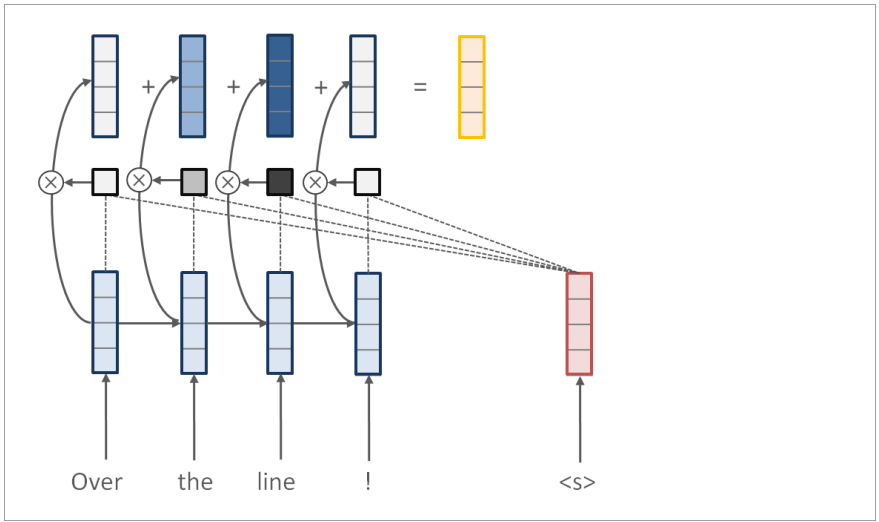
Seq2Seq+



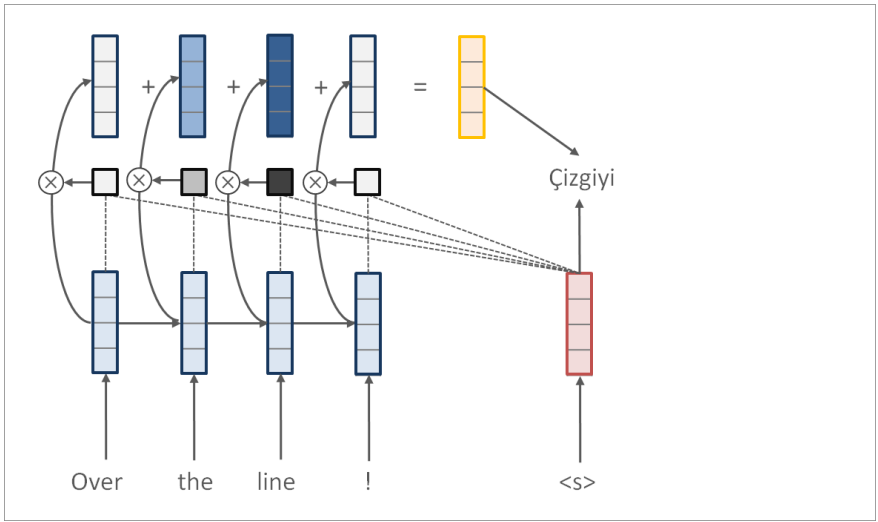
Seq2Seq+



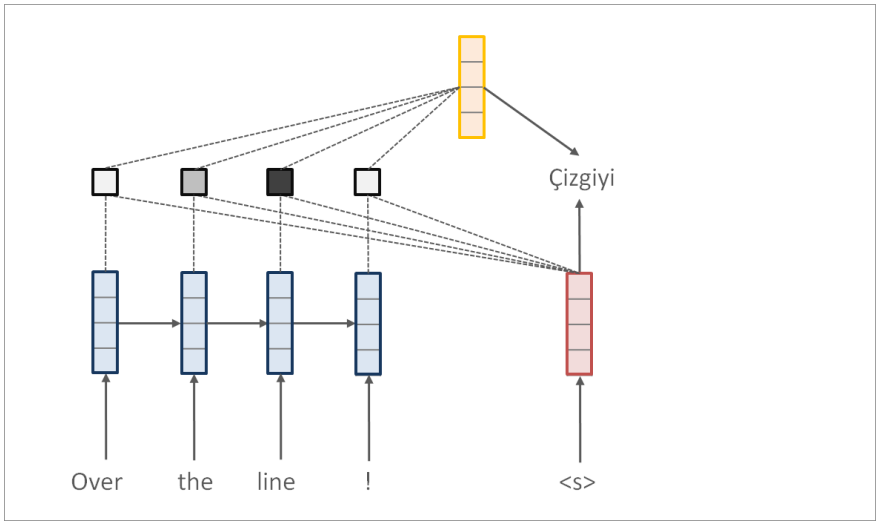
Seq2Seq+



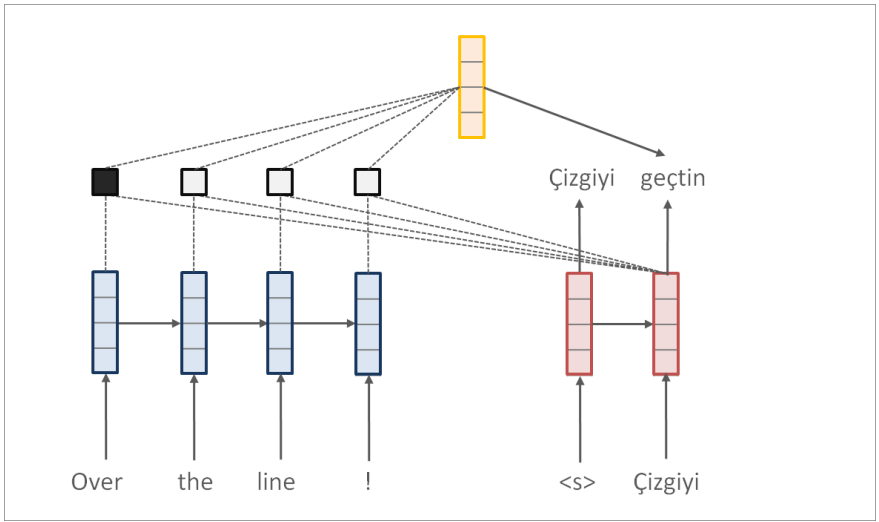
Seq2Seq+



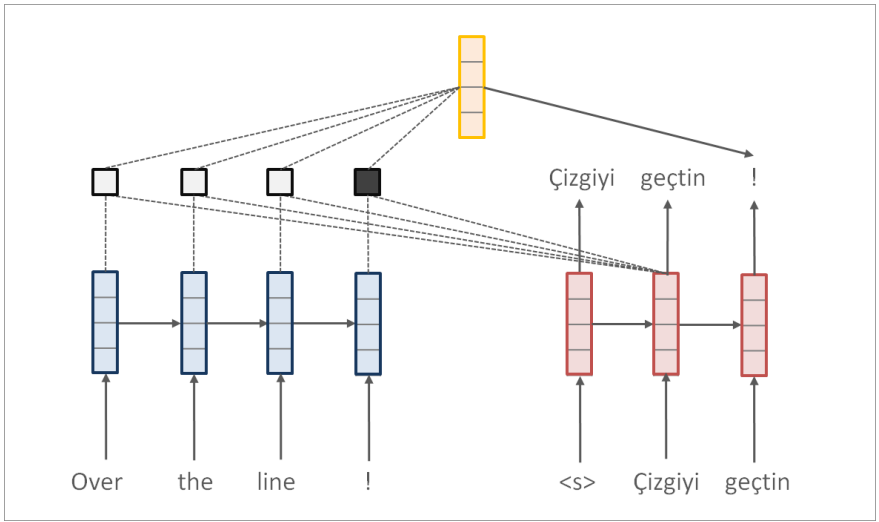
Seq2Seq+



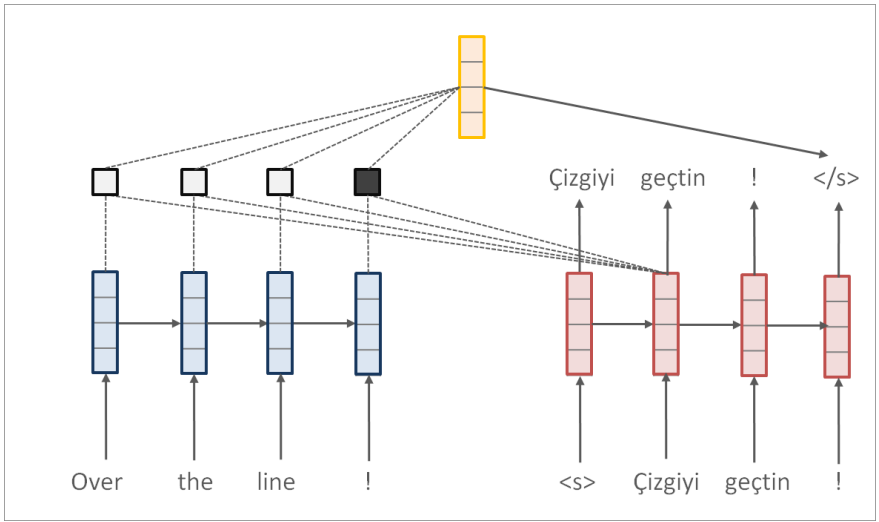
Seq2Seq+



Seq2Seq+



Seq2Seq+

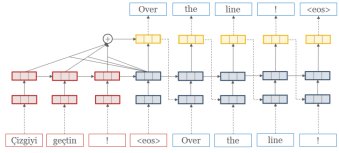


OpenNMT

(Klein et al, 2017)

Home

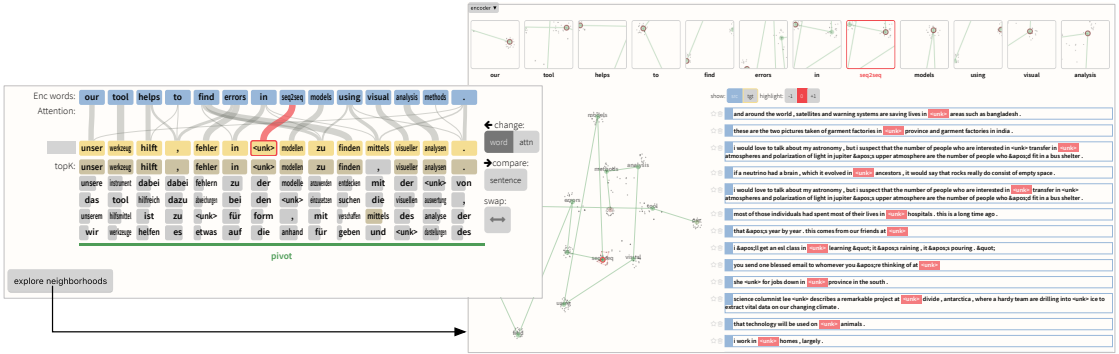
OpenNMT is a industrial-strength, open-source (MIT) neural machine translation system utilizing the Torch/PyTorch mathematical toolkit.



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)



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 ....

1 Introduction: Data-Driven Text Generation

2 Latent-Variable Generation

3 Work 1: Learning Neural Templates

4 Work 2: Learning Alignments

5 Intro

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.

Research Direction: Deep Latent-Variable Models for NLP

Goal: Expose specific choices as explicit latent variables.

Latent-Variable Model Basics

Latent variable models give us a joint distribution

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

- x is our observed data
- z is a collection of latent variables
- θ are the deterministic parameters of the model, such as the neural network parameters
- Data consists of N i.i.d samples,

$$p(x^{(1:N)}, z^{(1:N)}; \theta) = \prod_{n=1}^{N} p(x^{(n)} \mid z^{(n)}; \theta) p(z^{(n)}; \theta).$$

Latent-Variable Model Basics

Latent variable models give us a joint distribution

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

- x is our observed data
- z is a collection of latent variables
- θ are the deterministic parameters of the model, such as the neural network parameters
- Data consists of N i.i.d samples,

$$p(x^{(1:N)}, z^{(1:N)}; \theta) = \prod_{n=1}^{N} p(x^{(n)} \mid z^{(n)}; \theta) p(z^{(n)}; \theta).$$

Latent-Variable Model Basics

Latent variable models give us a joint distribution

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

- x is our observed data
- z is a collection of latent variables
- θ are the deterministic parameters of the model, such as the neural network parameters
- Data consists of N i.i.d samples,

$$p(x^{(1:N)}, z^{(1:N)}; \theta) = \prod_{n=1}^{N} p(x^{(n)} \mid z^{(n)}; \theta) p(z^{(n)}; \theta).$$

Posterior Inference

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

$$p(z \mid x; \theta) = \frac{p(x, z; \theta)}{p(x; \theta)}.$$

Why?

- z will often represent interesting information about our data.
- Intuition: if I know likely $z^{(n)}$ for $x^{(n)}$, I can learn by maximizing $p(x^{(n)} \mid z^{(n)}; \theta)$.

Posterior Inference

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

$$p(z \,|\, x;\, \theta) = \frac{p(x, z;\, \theta)}{p(x;\, \theta)}.$$

Why?

- z will often represent interesting information about our data.
- Intuition: if I know likely $z^{(n)}$ for $x^{(n)}$, I can learn by maximizing $p(x^{(n)} \,|\, z^{(n)};\, \theta)$.

Example: Copy-Attention

(Gu et al, 2016) (Gulcehre et al, 2016)

Let z be a binary latent variable.

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

Inference:

Pointer-generator model + coverage summary

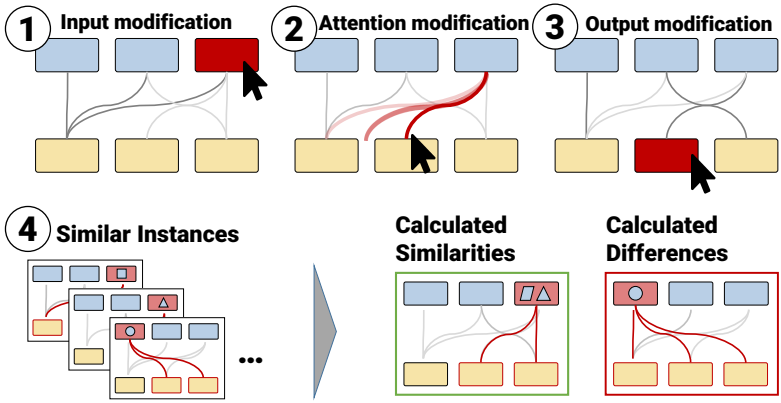
francis *saili* has signed a two-year deal to join munster later this year .
the 24-year-old was part of the new zealand under-20 side that won the junior world championship in italy in 2011 .
saili 's signature is something of a coup for munster and head coach anthony foley .

(See et al, 2017)

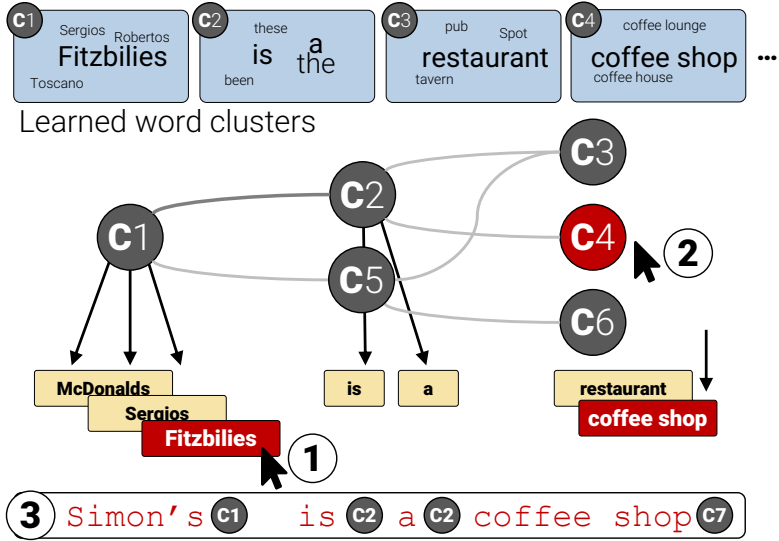
Latent Variable Models for Generation

- 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: Latent Alignment and Variational Attention



Approach 2: Learning Neural Templates



1 Introduction: Data-Driven Text Generation

2 Latent-Variable Generation

3 Work 1: Learning Neural Templates

4 Work 2: Learning Alignments

5 Intro

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 Miami 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. (botany)	Park.-Rhodes

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

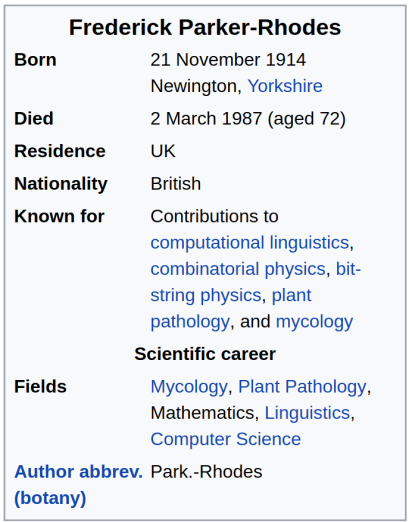
Data-to-Text Generation

[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."

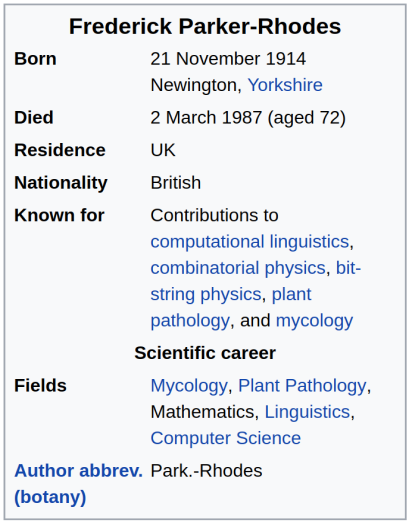
Argument for Templates #1: Interpretability



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. (botany)	Park.-Rhodes

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. (botany)	Park.-Rhodes

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

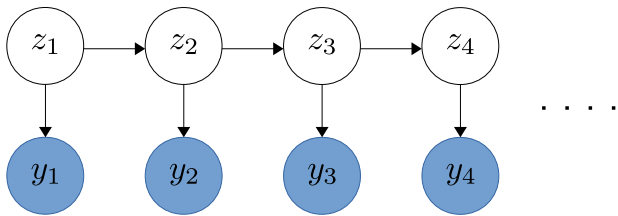
[Gales and Young, 1993; Ostendorf et al., 1996]

- 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]

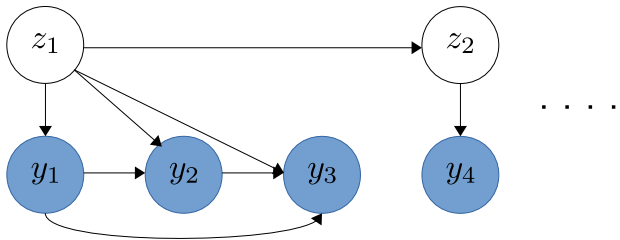
- Give a joint distribution over observations $y_{1:T}$ and discrete latents $z_{1:S}$
  - Like HMMs, but observations can last multiple time-steps:



Hidden Semi-Markov Models

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

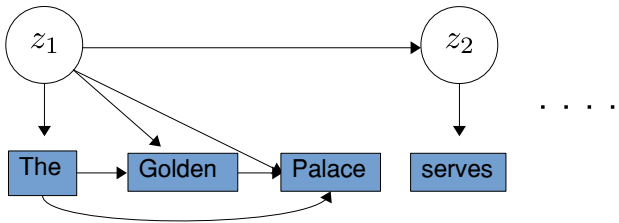
- Give a joint distribution over observations $y_{1:T}$ and discrete latents $z_{1:S}$
  - Like HMMs, but observations can last multiple time-steps:



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:

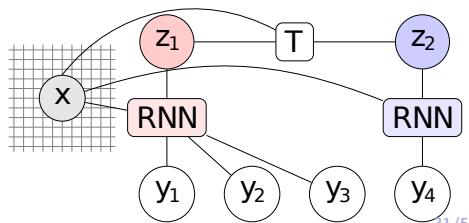


Upshot: HSMMs give us typed segmentations.

A Conditional (Neural) HSMM

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

- 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 $\ln p(y_{1:T} \mid x) = \ln \sum_z p(y_{1:T}, z \mid x)$
  - 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 $\arg\max_{y,z} p(y, z \mid x)$
- Instead, we'll first extract "templates":

  1) Viterbi-segment the training data

Generation

- Given an input x, we could generate by approximating $\arg\max_{y,z} p(y, z \mid x)$
- 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[£20-25]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 $\arg\max_{y,z} p(y, z \mid x)$
- 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[£20-25]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, ...

What's Good about these "Templates"?

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

1. | The Eagle
The Golden Curry
Zizzi
… | provides
providing
serves
… | Indian
Chinese
English
… | food
cuisine
Food
… | in the
with a
and has a
… | high
moderate
average
… | price range
customer rating
rating
… | . | It is
They are
It's
… | near
located in the
located near
… |
   | riverside
city centre
Cafe Sicilia
… | . | Its customer rating is
It has a
The price range is
… | 1 out of 5
average
high
… | .

2. | Located near
Located in the
Near
… | The Portland Arms
riverside
city centre
… | is an
is a family friendly
there is a
… | Italian
fast food
French
… | restaurant called
place called
restaurant named
… | The Waterman
Cocum
Loch Fyne
… | .

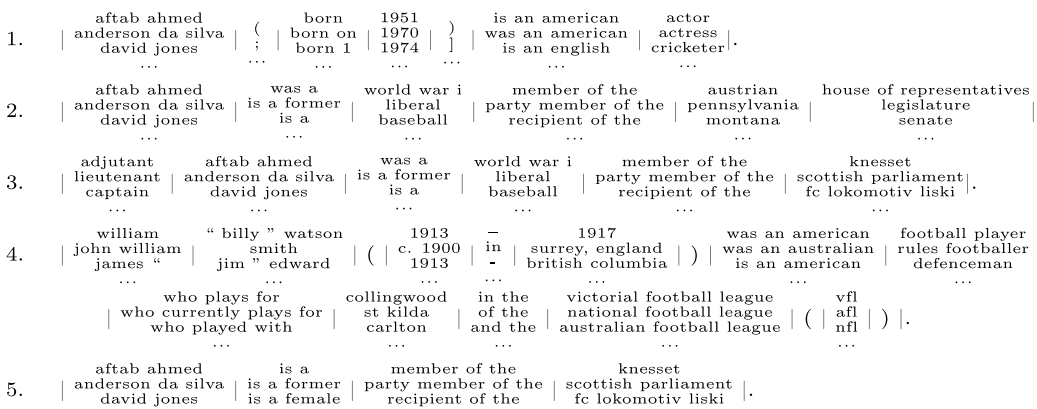
3. | A
An
A family friendly
… | Italian
fast food
French
… | restaurant
pub
coffee shop
… | is
called
named
… | The Waterman
Cocum
Loch Fyne
… | .

4. | Located near
Located in the
Near
… | The Portland Arms
riverside
city centre
… | , | The Eagle
The Golden Curry
Zizzi
… | is a
is a family friendly
is an
… | cheap
family-friendly
family friendly
… | Italian
fast food
French
… | restaurant
pub
coffee shop
… | .

5. | A
An
A family friendly
… | Italian
fast food
French
… | restaurant
pub
coffee shop
… | near
located in the
located near
… | riverside
city centre
Cafe Sicilia
… | is
called
named
… | The Waterman
Cocum
Loch Fyne
… | .

What's Good about these "Templates"?

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



What's Good about these "Templates"?

2) We can use them to control generation:

- Select a template $z^{(i)} = z_1^{(i)}, \ldots, z_S^{(i)}$
- Generate by computing $\arg\max_y p(y, z = z^{(i)} \mid x)$
- 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” $z^{(i)} = z_1^{(i)}, \ldots, z_S^{(i)}$

3) Given a new input x, generate by finding $\arg\max_y p(y, z^{(i)} \mid x)$ 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

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

Travellers Rest Beefeater$_{55}$ is a$_{59}$ 3 star$_{43}$ restaurant$_{11}$ located near$_{25}$ Raja Indian Cuisine$_{40}$.$_{53}$

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 Beefeater$_{55}$ is a$_{59}$ 3 star$_{43}$ restaurant$_{11}$ located near$_{25}$ Raja Indian Cuisine$_{40}$.$_{53}$

Near$_{31}$ riverside$_{29}$,$_{44}$ Travellers Rest Beefeater$_{55}$ serves$_{3}$ 3 star$_{50}$ 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 Beefeater$_{55}$ is a$_{59}$ 3 star$_{43}$ restaurant$_{11}$ located near$_{25}$ Raja Indian Cuisine$_{40}$.$_{53}$

Near$_{31}$ riverside$_{29}$,$_{44}$ Travellers Rest Beefeater$_{55}$ serves$_{3}$ 3 star$_{50}$ food$_{1}$.$_{2}$

Travellers Rest Beefeater$_{55}$ is a$_{59}$ restaurant$_{12}$ providing$_{3}$ riverside$_{50}$ food$_{1}$ and has a$_{17}$ 3 out of 5$_{26}$ customer rating$_{16}$.$_{2}$ It is$_{8}$ near$_{25}$ 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 Beefeater$_{55}$ is a$_{59}$ 3 star$_{43}$ restaurant$_{11}$ located near$_{25}$ Raja Indian Cuisine$_{40}$.$_{53}$

Near$_{31}$ riverside$_{29}$,$_{44}$ Travellers Rest Beefeater$_{55}$ serves$_{3}$ 3 star$_{50}$ food$_{1}$.$_{2}$

Travellers Rest Beefeater$_{55}$ is a$_{59}$ restaurant$_{12}$ providing$_{3}$ riverside$_{50}$ food$_{1}$ and has a$_{17}$ 3 out of 5$_{26}$ customer rating$_{16}$.$_{2}$ It is$_{8}$ near$_{25}$ Raja Indian Cuisine$_{40}$.$_{53}$

Travellers Rest Beefeater$_{55}$ is a$_{59}$ place to eat$_{12}$ located near$_{25}$ Raja Indian Cuisine$_{40}$.$_{53}$

Travellers Rest Beefeater$_{55}$ is a$_{59}$ 3 out of 5$_{5}$ rated$_{32}$ riverside$_{43}$ restaurant$_{11}$ near$_{25}$ 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 deacon$_{42}$ ($_{30}$ born$_{44}$ 23 january 1906$_{11}$)$_{22}$ was a$_{14}$ scottish$_{8}$ football$_{19}$ forward$_{24}$.$_{43}$

yang sung-chul$_{42}$ ($_{30}$ november 20, 1939$_{11}$ in$_{21}$ gokseong county, jeollanam-do$_{39}$)$_{22}$ is a$_{14}$ south korean$_{8}$ political scientist$_{24}$.$_{43}$

1 Introduction: Data-Driven Text Generation

2 Latent-Variable Generation

3 Work 1: Learning Neural Templates

4 Work 2: Learning Alignments

5 Intro

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 chiapas states hardest hit. authorities say 700,000 affected ...

Six Challenges for NMT (Koehn and Knowles 2017)

1. Out of domain generalization
2. **Sample complexity**
3. Rare words
4. Long sentences
5. **The alignments learned by soft attention may not be interpreted as word alignments**
6. Beam search degradation for large beams

1 Introduction: Data-Driven Text Generation

2 Latent-Variable Generation

3 Work 1: Learning Neural Templates

4 Work 2: Learning Alignments

5 Intro

Attention versus Alignment

- Attention is motivated as an "alignment" module:
  - makes the model's behavior interpretable
  - can be used for other prediction tasks
- Soft Attention is a *deterministic* from a neural network,
- We contrast this with alignment which acts as a random variable.

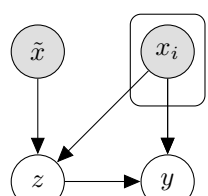
Latent Alignment: Motivation

If attention works so well, why study alignment?

- A latent variable approach facilitates composibility 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

- $x = x_1, \cdots, x_T$: 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

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

- Training Objective (maximizing marginal log-likelihood)

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

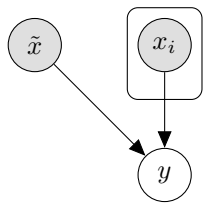
- Direct optimization is computationally expensive
  - Discrete $z \sim \mathcal{D}$: $O(T)$ additional runtime
  - Continuous $z \sim \mathcal{D}$: intractable

Workaround 1: Soft Attention

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

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

- The corresponding graphical model is



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

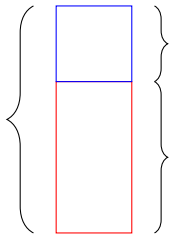
$$\log \mathbb{E}_z[f(x, z; \theta)_{\hat{y}}] \geq \mathbb{E}_z \log[f(x, z; \theta)_{\hat{y}}]$$

- The use of the prior in the expectation may result in a poor bound

Marginal Likelihood: Variational Decomposition

For any[1] distribution $q(z)$ over z,

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



posterior gap

ELBO (evidence lower bound)

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

[1] Technical condition: $\mathsf{supp}(q(z)) \subset \mathsf{supp}(p(z \mid x; \theta))$

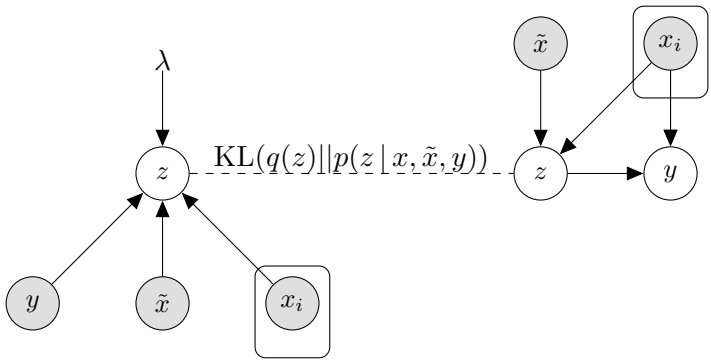
Proposal: Variational Attention

- Learn model and q to maximize the following lower bound

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

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

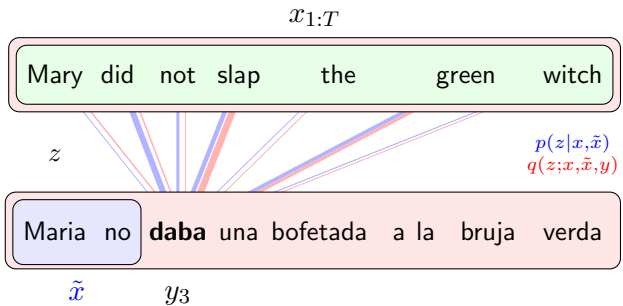
Example Form of q: Amortized Parameterization



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

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

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

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

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 $f(x, \mathbb{E}_{z' \sim q(z)}[z'])$

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	$\log p(y \mid \mathbb{E}[z])$	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	$\mathbb{E}_q[\log p] - \mathrm{KL}$	Sample	7.58	30.05
Var Attn	$\mathbb{E}_q[\log p] - \mathrm{KL}$	Enum	6.08	**33.69**
Var Attn	$\mathbb{E}_q[\log p] - \mathrm{KL}$	Sample	6.17	**33.30**

Preliminary Results (Low Data Settings)

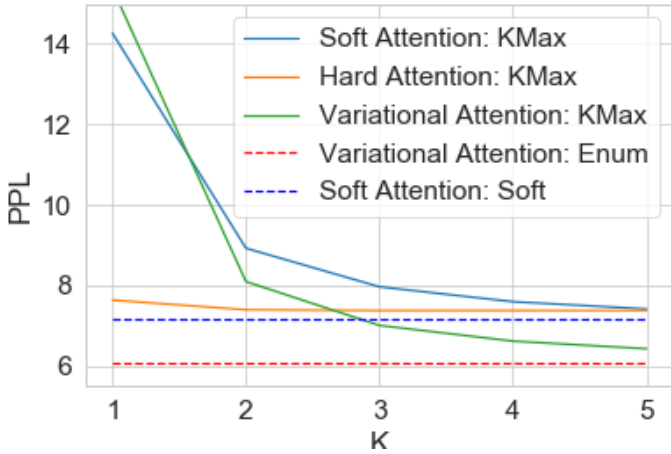
Training Size	10k	25k	50k	Full (160k)
Soft Attn	12.10	22.88	26.65	32.77
Marginal Likelihood	16.79	23.44	26.99	33.29
Var Attn Enum	14.90	23.50	27.26	33.69
Var Attn Sample	12.20	23.35	27.87	33.30

Results (VQA)

Model	Objective	Exp	NLL	Eval
Soft Attn	$\log p(y \mid \mathbb{E}[z])$	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	$\mathbb{E}_q[\log p] - \mathrm{KL}$	Enum	1.68	58.44
Var Attn	$\mathbb{E}_q[\log p] - \mathrm{KL}$	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


Soft Attention: KMax
Hard Attention: KMax
Variational Attention: KMax
Variational Attention: Enum
Soft Attention: Soft
PPL
14
12
10
8
6
1
2
3
4
5
K

Discussion

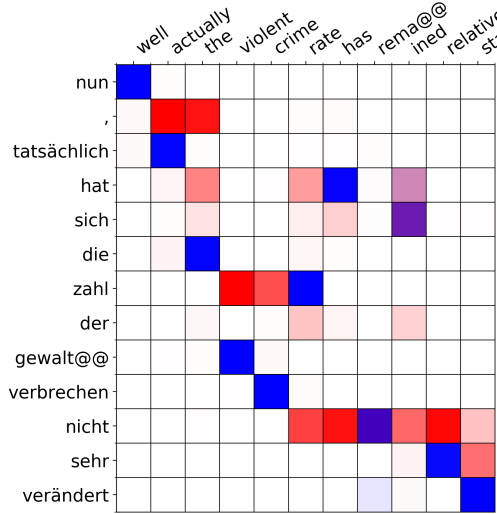
- Soft attention underperforms latent attention
- Variational attention achieves similar performance to maximizing the marginal likelihood exactly despite optimizing a lower bound
- Proposed soft attention as a computationally cheap baseline for reducing the variance of the REINFORCE gradient estimator

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

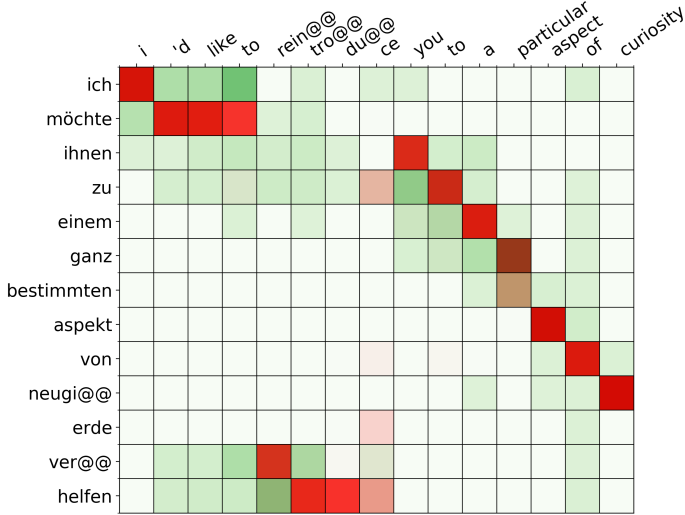
- Gumbel-Softmax is a viable alternative
- RWS incurs higher memory cost

Example Alignments



Red: prior; blue: posterior.

Example Alignments



Red: prior; green: soft attention.