Controlling Text Generation

Alexander Rush (and Sam Wiseman)
Harvard / Cornell Tech

GANocracy

Outline

- **Background: Text Generation**
- Latent-Variable Generation
- Learning Neural Templates

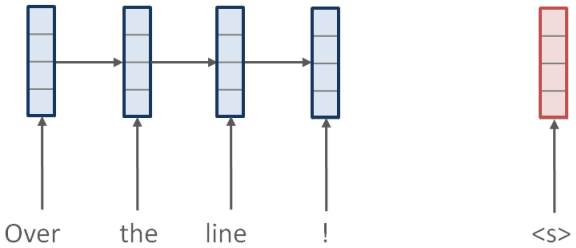
Machine Learning for Text Generation

$$y^*_{1:T} = \arg\max_{y_{1:T}} p_\theta(y_{1:T} \mid x)$$

- Input x, *what to talk about*
- Possible output text $y_{1:T}$, *how to say it*
- Scoring function p_θ, with parameters θ learned from data

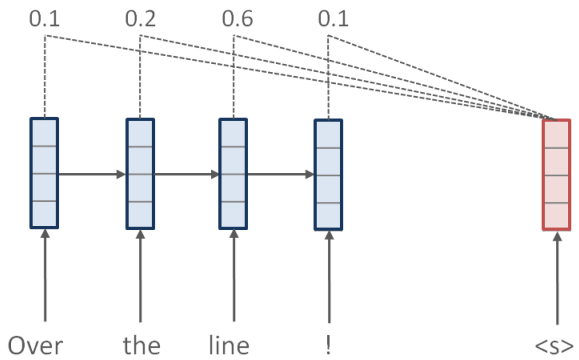
Attention-Based Decoding

$p_\theta(y_{1:T} \mid x)$



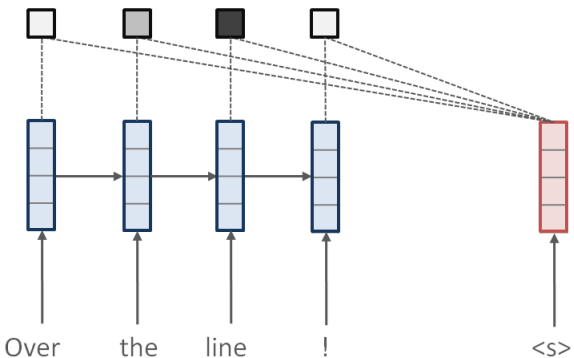
Attention-Based Decoding

$p_\theta(y_{1:T} \mid x)$



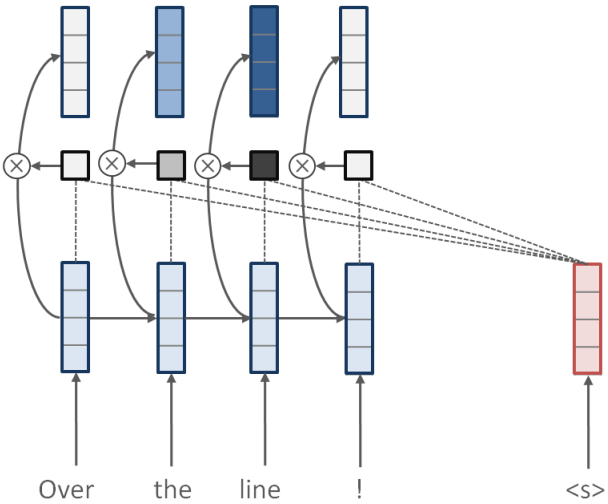
Attention-Based Decoding

$p_\theta(y_{1:T} \mid x)$



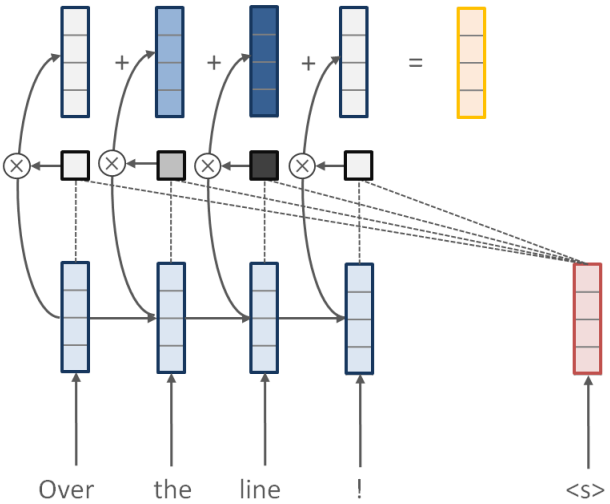
Attention-Based Decoding

$p_\theta(y_{1:T} \mid x)$



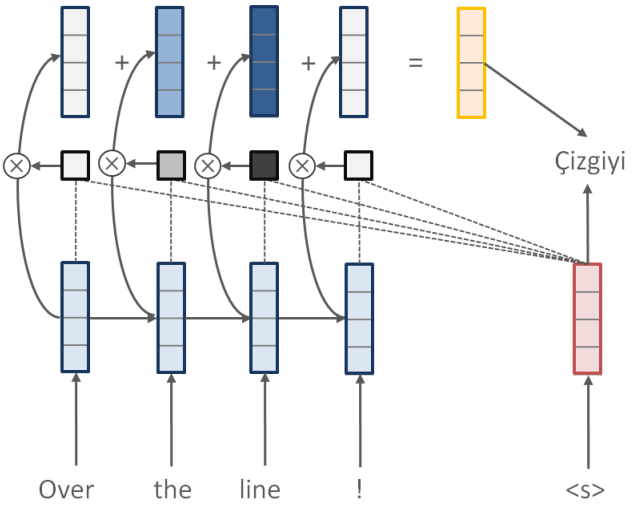
Attention-Based Decoding

$p_\theta(y_{1:T} \mid x)$



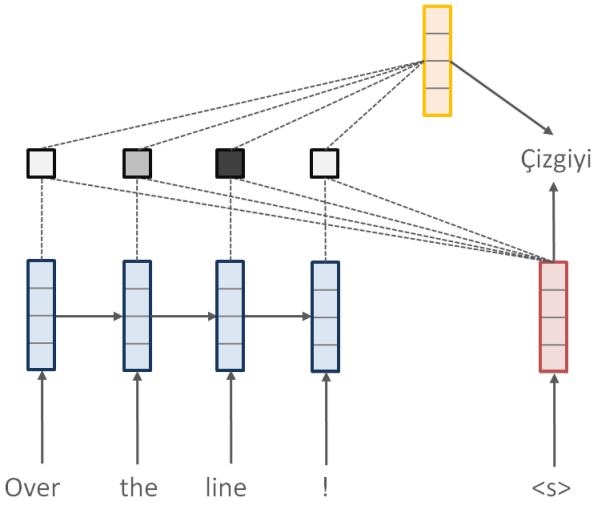
Attention-Based Decoding

$p_\theta(y_{1:T} \mid x)$



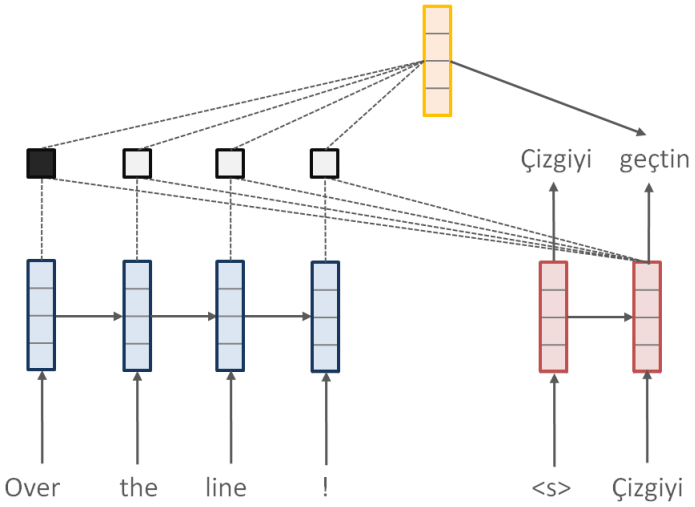
Attention-Based Decoding

$p_\theta(y_{1:T} \mid x)$



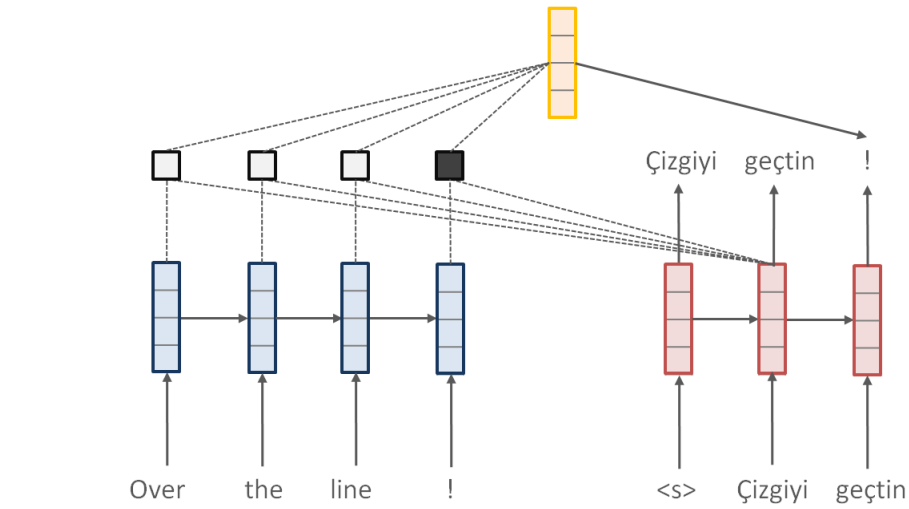
Attention-Based Decoding

$p_\theta(y_{1:T} \mid x)$



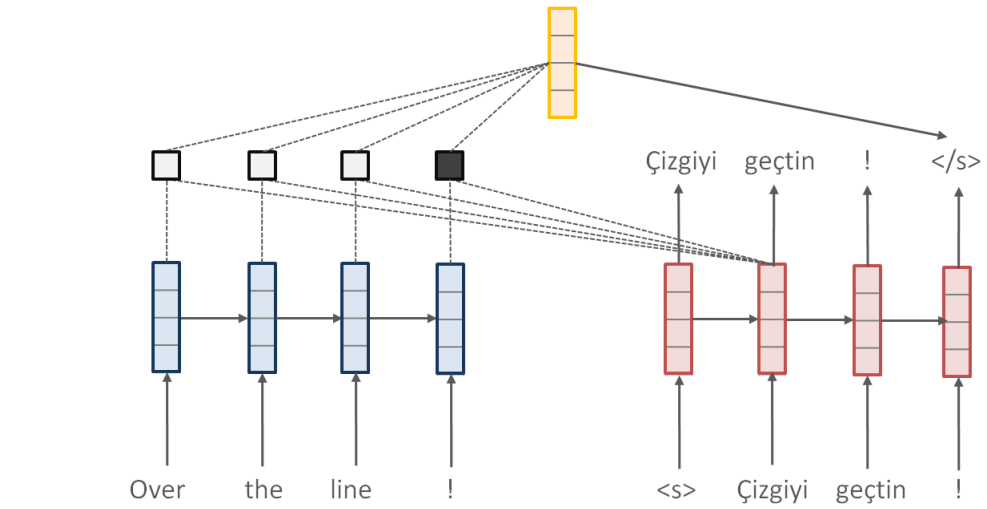
Attention-Based Decoding

$p_\theta(y_{1:T} \mid x)$



Attention-Based Decoding

$p_\theta(y_{1:T} \mid x)$



Talk about Text

London, England (reuters) – Harry Potter star Daniel Radcliffe gains access to a reported $20 million fortune as he turns 18 on monday, but he insists the money won't cast a spell on him. Daniel Radcliffe as harry potter in "Harry Potter and the Order of the Phoenix" to the disappointment of gossip columnists around the world , the young actor says he has no plans to fritter his cash away on fast cars , drink and celebrity parties . " i do n't plan to be one of those people who , as soon as they turn 18 , suddenly buy themselves a massive sports car collection or something similar , " he told an australian interviewer earlier this month . " i do n't think i 'll be particularly extravagant " . " the things i like buying are things that cost about 10 pounds – books and cds and dvds . " at 18 , radcliffe will be able to gamble in a casino , buy a drink in a pub or see the horror film " hostel : part ii , " currently six places below his number one movie on the uk box office chart . details of how he 'll mark his landmark birthday are under wraps . his agent and publicist had no comment on his plans . " i 'll definitely have some sort of party , " he said in an interview ...

Harry Potter star Daniel Radcliffe gets $20m fortune as he turns 18 monday. Young actor says he has no plans to fritter his cash away. Radcliffe's earnings from first five potter films have been held in trust fund.

Talk about Diagrams

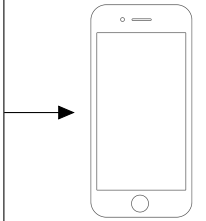
$$\mathcal{K}^L(\sigma = 2) = \left(\begin{array}{cc} -\frac{d^2}{dx^2} + 4 - \frac{3}{\cosh^2 x} & \frac{3}{\cosh^2 x} \\ \frac{3}{\cosh^2 x} & -\frac{d^2}{dx^2} + 4 - \frac{3}{\cosh^2 x} \end{array} \right) \quad ,$$

```
 { \cal K } ^ { L } ( \sigma = 2 ) = \left( \begin{array}
{ c c } { - \frac { d ^ { 2 } } { d x ^ { 2 } } + 4 - \frac
 { 3 } { \operatorname { c o s h } ^ { 2 } x } } \& { \frac
 { 3 } { d x ^ { 2 } } }  { \frac { 3 } { \operatorname
 { c o s h } ^ { 2 } x } } \& { - \frac { d ^ { 2 } }
{ d x ^ { 2 } } + 4 - \frac { 3 } { \operatorname { c o s h }
^ { 2 } x } }  \end{array} \right) \qquad
```

Talk about Data

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

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



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 he amassed eight points, 12 rebounds and one blocks on 4 - of - 12 shooting ...

Outline

- Background: Text Generation
- **Latent-Variable Generation**
- Learning Neural Templates

Why DL People Say I Need GANs

- They produce *awesome* unconditional samples.
  - What if auto-regressive models are far superior for text?

Why DL People Say I Need GANs

- They produce *awesome* unconditional samples.
  - What if auto-regressive models are far superior for text?
- They model latent variables.
  - What's the point if I can't do posterior inference?

Why DL People Say I Need GANs

- They produce *awesome* unconditional samples.
  - What if auto-regressive models are far superior for text?
- They model latent variables.
  - What's the point if I can't do posterior inference?
- They allow for interpolations.
  - Should I expect language to be continuous?

What I Need From Generative Models

Structure induction from latent variables z.

$$p_\theta(y, z \mid x)$$

- x, y as before, *what to talk about, how to say it*
- z is a collection of problem-specific discrete latent variables, *why we said it that way*

What I Need From Generative Models

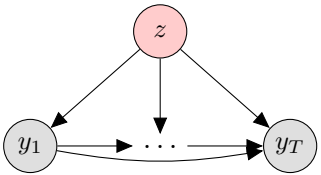
Structure induction from latent variables z.

$$p_\theta(y, z \mid x)$$

- x, y as before, *what to talk about, how to say it*
- z is a collection of problem-specific discrete latent variables, *why we said it that way*

?

Motivating Model: Clustering



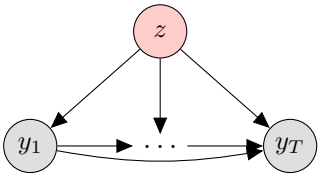
The film is the first from ... $z = 1$

Allen shot four-for-nine ... $z = 2$

In the last poll Ericson led ... $z = 3$

1. Draw cluster $z \in \{1, \ldots, Z\}$.
2. Draw word sequence $y_{1:T}$ from decoder RNN z.

Motivating Model: Clustering



The film is the first from ... $z = 1$

Allen shot four-for-nine ... $z = 2$

In the last poll Ericson led ... $z = 3$

1. Draw cluster $z \in \{1, \ldots, Z\}$.
2. Draw word sequence $y_{1:T}$ from decoder RNN z.

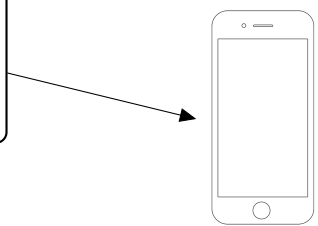
Outline

- Background: Text Generation
- Latent-Variable Generation
- **Learning Neural Templates**

Talk about Data

x

Fitzbillies	
type	coffee shop
price	$<$ £20
food	Chinese
rating	3/5
area	city centre



p_θ

Talk about Data

x

Fitzbillies	
type	coffee shop
price	$<$ £20
food	Chinese
rating	3/5
area	city centre

p_θ

$y_{1:T}$

Fitzbillies is a coffee shop providing Chinese food in the moderate price range . It is located in the city centre . Its customer rating is 3 out of 5 .

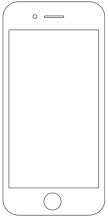
Talking About Data

x

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

p_θ



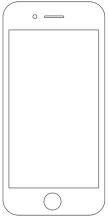
Talking About Data

x

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

p_θ



$y_{1:T}$

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

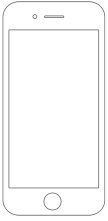
Talking About Data

x

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

p_θ



$y^*_{1:T}$

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

Talking About Data

x

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

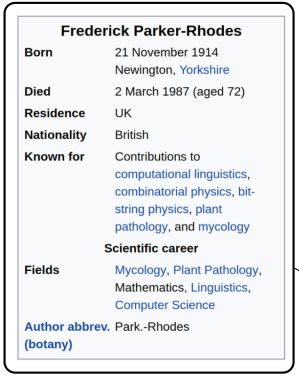
p_θ

$z_{1:T}$

___ (born ___) was a ___ ___ , who lived in the ___ . He was known for contributions to ___ .

Talking About Data

x



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

$z_{1:T}$

___ (born ___) was a ___ ___ , who lived in the ___ . He was known for contributions to ___ .

p_θ

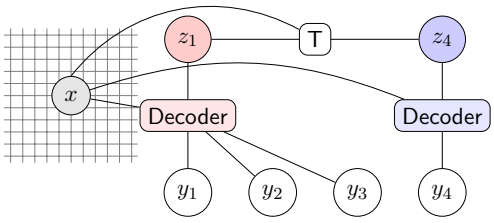
$y^*_{1:T}$

Frederick Parker-Rhodes (born 21 November 1914) was a English mycologist who lived in the UK. He was known for contributions to plant pathology.

Model: A Deep Hidden Semi-Markov Model

Hidden Semi-Markov Model

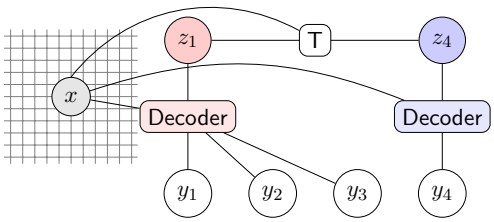
Distribution: Encoder-Decoder, specialized per cluster $\{1, \ldots, Z\}$.



Model: A Deep Hidden Semi-Markov Model

Hidden Semi-Markov Model

Distribution: Encoder-Decoder, specialized per cluster $\{1, \ldots, Z\}$.



Probabilistic Model $\Rightarrow$ Templates

(Step 1) Train (Step 2) Match (Step 3) Extract

Step 1: Training HSMM

Training requires summing over clusters and segmentation of deep model.

$$\mathcal{L}(\theta) = \log \mathbb{E}_{z_{1:T}} p_\theta(\hat{y}_{1:T} \mid z_{1:T},\ x) = \log \sum_{z_{1:T}} p_\theta(\hat{y}_{1:T}, z_{1:T} \mid x)$$

Step 1: Training HSMM

Training requires summing over clusters and segmentation of deep model.

$$\mathcal{L}(\theta) = \log \mathbb{E}_{z_{1:T}} p_\theta(\hat{y}_{1:T} \mid z_{1:T},\ x) = \log \sum_{z_{1:T}} p_\theta(\hat{y}_{1:T}, z_{1:T} \mid x)$$

Example

$\hat{y}_{1:T}$ = Frederick Parker-Rhodes was an English linguist, plant pathologist ...

$$\Downarrow \sum_{z_{1:T}} p_\theta(\hat{y}_{1:T}, z_{1:T} \mid x)$$

Frederick Parker-Rhodes | was an English | linguist , | plant pathologist ...

Frederick | Parker-Rhodes | was an | English linguist , | plant pathologist ...

Frederick | Parker-Rhodes was an English linguist , | linguist , | plant pathologist ...

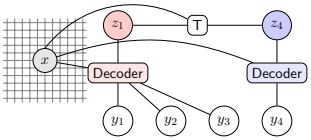
Step 1: Technical Methodology

Training is end-to-end, i.e. clusters and segmentation are learned simultaneously with encoder-decoder model on GPU.

- Backpropagation through dynamic programming.
- Parameters are trained by exactly marginalizing over segmentations, equivalent to expectation-maximization.
- Utilize HSMM backward algorithm within standard training.

Step 2: Template Assignment

Finding best/Viterbi cluster sequences for each training sentence.



$$z^*_{1:T} = \arg\max_{z_{1:T}} p_\theta(y_{1:T}, z_{1:T} \mid x)$$

Example

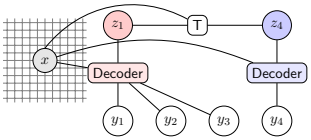
Frederick Parker-Rhodes was an English linguist, plant pathologist

$$\Downarrow \arg\max_{z_{1:T}}$$

Frederick Parker-Rhodes | was an English | linguist , | plant pathologist | ...

Step 2: Template Assignment

Finding best/Viterbi cluster sequences for each training sentence.



$$z^*_{1:T} = \arg\max_{z_{1:T}} p_\theta(y_{1:T}, z_{1:T} \mid x)$$

Example

Frederick Parker-Rhodes was an English linguist, plant pathologist

$$\Downarrow \arg\max_{z_{1:T}}$$

Frederick Parker-Rhodes | was an English | linguist , | plant pathologist | ...

Step 3: Template Extraction

Find identical cluster sequences $z_{1:T}$ that occur most often in training data.

Frederick Parker-Rhodes was an English linguist, plant pathologist ...

Bill Jones was an American professor, and well-known author ...

⋮

$\Downarrow \arg\max_{z_{1:T}}$

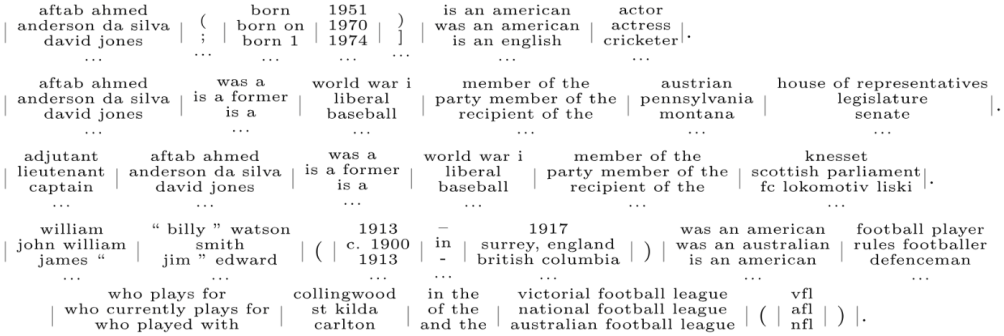
Frederick Parker-Rhodes | was an English | linguist , | plant pathologist ...

Bill Jones | was an American | professor , | and well-known author ...

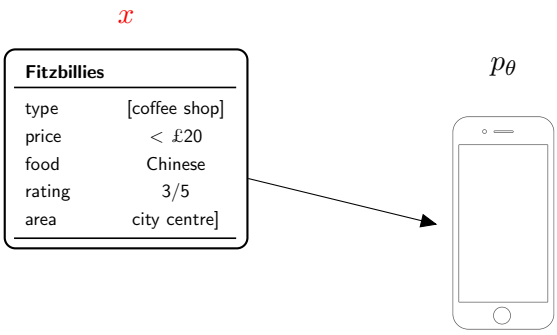
⋮

Example Templates: Wikipedia

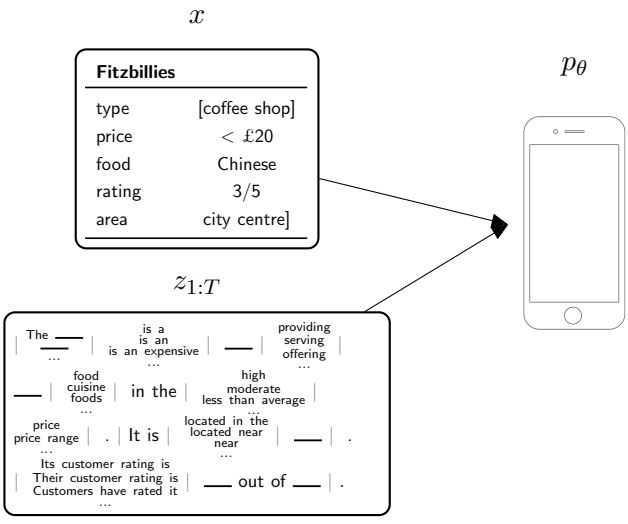
Example common extracted “templates”.



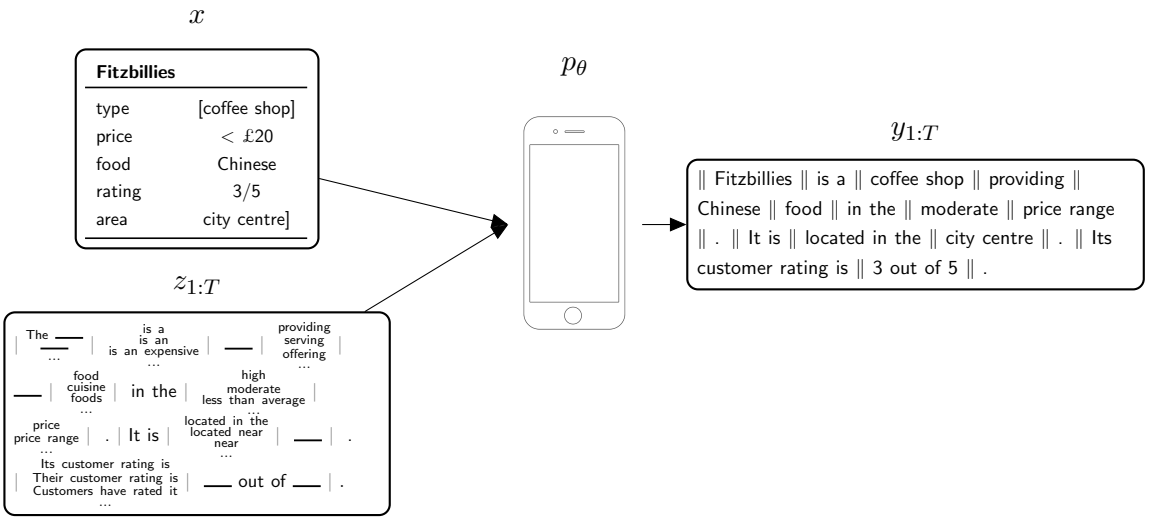
Neural Template Generation Approach



Neural Template Generation Approach



Neural Template Generation Approach



Interpretable Output

kenny warren

name: kenny warren, **birth date:** 1 april 1946,
birth name: kenneth warren deutscher, **birth place:** brooklyn, new york,
occupation: ventriloquist, comedian, author,
notable work: book - the revival of ventriloquism in america

1. kenny warren deutscher (april 1, 1946) is an american ventriloquist.
2. kenny warren deutscher (april 1, 1946 , brooklyn,) is an american ventriloquist.
3. kenny warren deutscher (april 1, 1946) is an american ventriloquist, best known for his the revival of ventriloquism.
4. "kenny" warren is an american ventriloquist.
5. kenneth warren "kenny" warren (born april 1, 1946) is an american ventriloquist, and author.

Controllable Style

The Golden Palace

name[The Golden Palace], type[coffee shop], food[Chinese],
priceRange[cheap] custRating[5 out of 5], area[city centre],

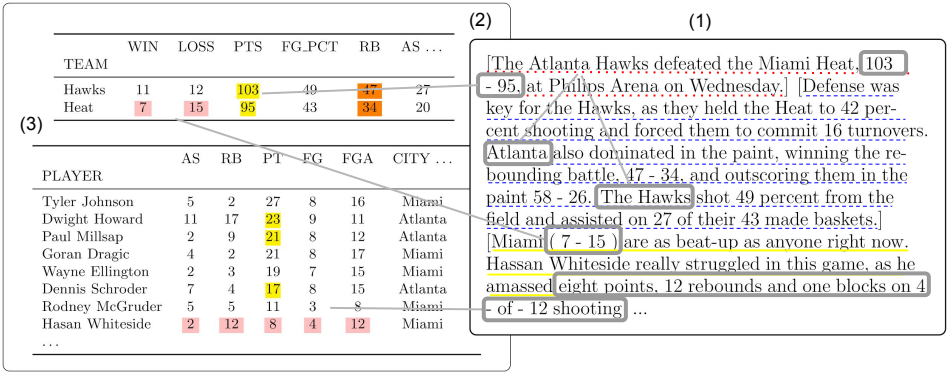
1. The Golden Palace is a coffee shop located in the city centre.
2. In the city centre is a cheap Chinese coffee shop called The Golden Palace.
3. The Golden Palace is a Chinese coffee shop.
4. The Golden Palace is a Chinese coffee shop with a customer rating of 5 out of 5.
5. The Golden Palace that serves Chinese food in the cheap price range. It is located in the city centre. Its customer rating is 5 out of 5.

Automatic Metrics

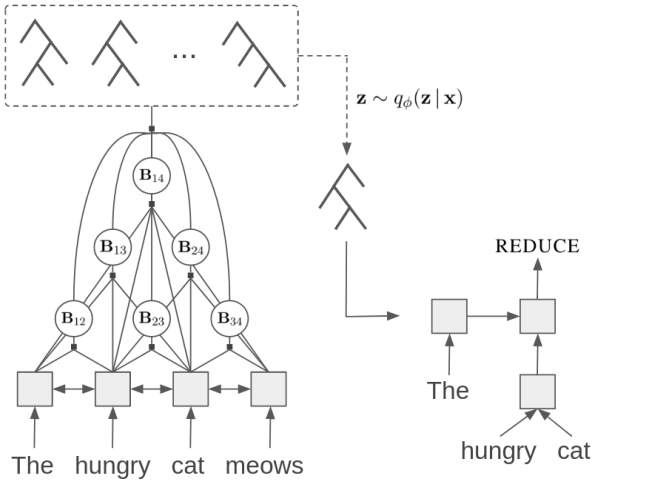
	Reviews (ROUGE)
Template	54.6
Neural Template	65.0
Best Model	68.5

	WikiBio (BLEU)
Template	19.8
Neural Template	34.7
Best Model	34.8

Future Work: Reasoning Systems for Long-Form Generation

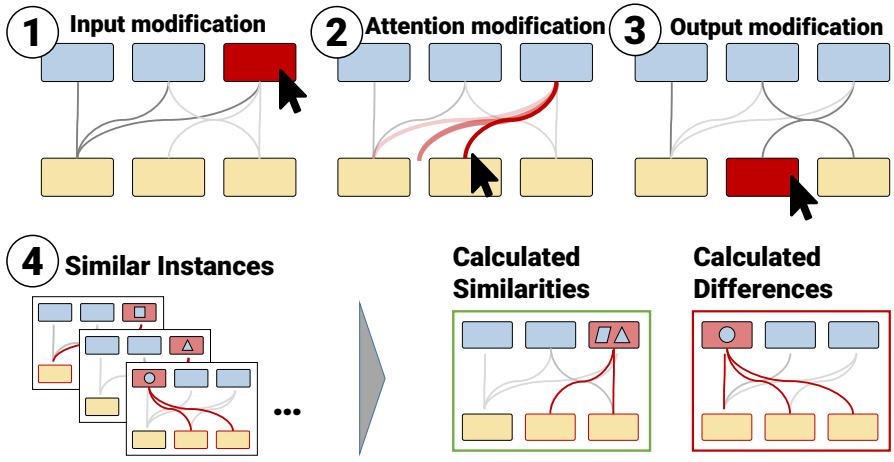


Future Work: Inducing Grammatical Structure



Inference Network $q_\phi(\mathbf{z} \mid \mathbf{x})$ Generative Model $p_\theta(\mathbf{x}, \mathbf{z})$

Future Work: Controllable Deep Learning for Generation



Thanks!

Yuntian Deng, Anssi Kanervisto, and Alexander M. Rush. 2016. What You Get Is What You See: A Visual Markup Decompiler. In *Arxiv*.

Yuntian Deng, Yoon Kim, Justin Chiu, Demi Guo, and Alexander Rush. 2018. Latent alignment and variational attention. In *Advances in Neural Information Processing Systems*, pages 9735–9747.

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M. Rush. 2017. Structured attention networks. abs/1702.00887.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. 2016. Character-Aware Neural Language Models. In *AAAI*.

Yoon Kim and Alexander M. Rush. 2016. Sequence-Level Knowledge Distillation. In *EMNLP*.

Yoon Kim, Sam Wiseman, Andrew C. Miller, David Sontag, and Alexander M. Rush. 2018. Semi-amortized variational autoencoders.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017.

Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, System Demonstrations*, pages 67–72.

Brandon Reagen, Udit Gupta, Robert Adolf, Michael M Mitzenmacher, Alexander M Rush, Gu-Yeon Wei, and David Brooks. 2017. Weightless: Lossy weight encoding for deep neural network compression. *arXiv preprint arXiv:1711.04686*.

Alexander Rush. 2018. The annotated transformer. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 52–60.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A Neural Attention Model for Abstractive Sentence Summarization. In *EMNLP*, September, pages 379–389.

Allen Schmaltz, Yoon Kim, Alexander M. Rush, and Stuart M. Shieber. 2016. Sentence-Level Grammatical Error Identification as Sequence-to-Sequence Correction. In *arxiv*.

Jean Senellart, Dakun Zhang, WANG Bo, Guillaume Klein, Jean-Pierre Ramatchandirin,

Josep Crego, and Alexander Rush. 2018. Opennmt system description for wnmt 2018: 800 words/sec on a single-core cpu. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 122–128.

Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, and Alexander M Rush. 2019. Seq2seq-v is: A visual debugging tool for sequence-to-sequence models. *IEEE transactions on visualization and computer graphics*, 25(1):353–363.

Hendrik Strobelt, Sebastian Gehrmann, Bernd Huber, Hanspeter Pfister, and Alexander M. Rush. 2016. Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks. In *Arxiv*.

Sam Wiseman, Alexander M. Rush, and Stuart M. Shieber. 2017a. Challenges in Data-to-Document Generation. In *EMNLP*.

Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. 2017b. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical*

Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, pages 2253–2263.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2018. Learning neural templates for text generation. *arXiv preprint arXiv:1808.10122*.