

Controlling Text Generation

Alexander Rush / HarvardNLP

Outline

Goal Controllable Generation

Outline

Goal Controllable Generation

- **Model and Background**
- Work 1: Generation
- Work 2: Attention
- Challenges: Text Generation and Deep Learning

Machine Learning for Text Generation

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

- Input $\textcolor{red}{x}$, what to talk about

Machine Learning for Text Generation

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

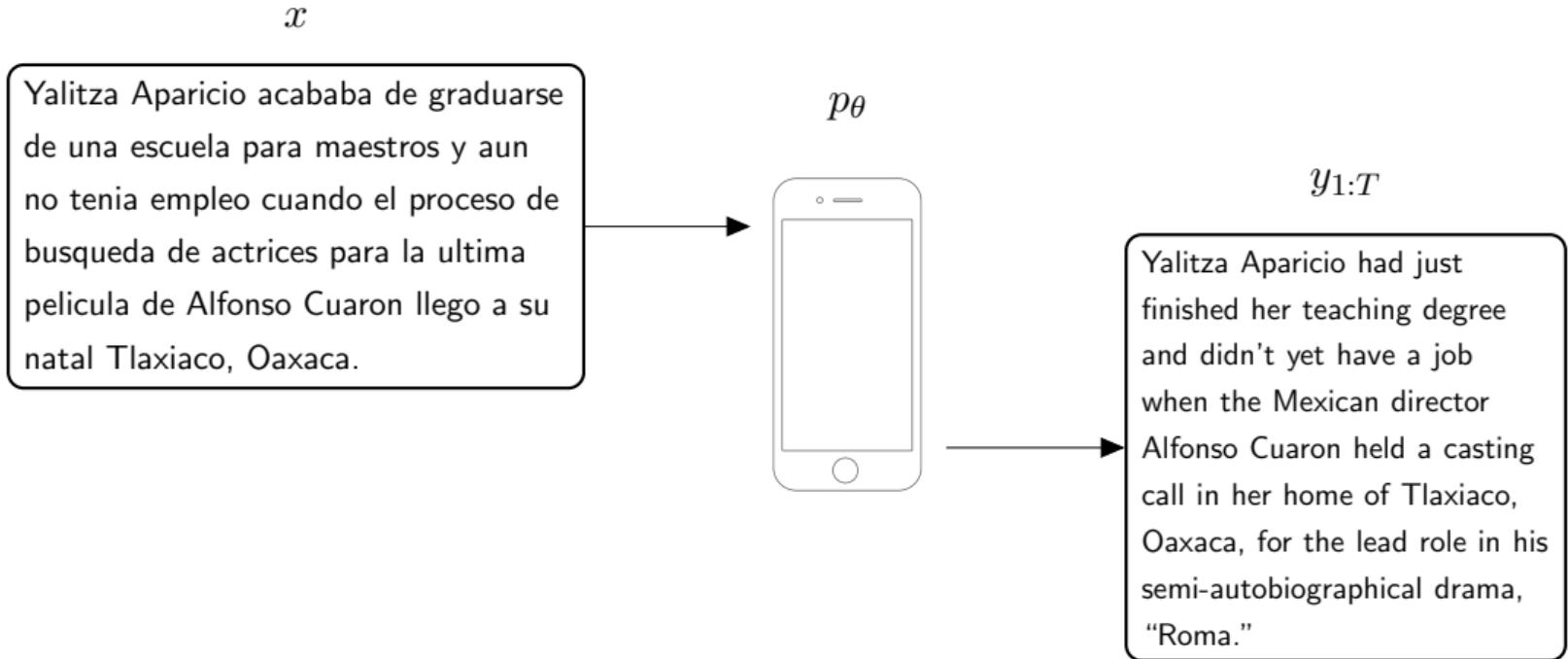
- Input x , *what to talk about*
- Possible output text $y_{1:T}$, *how to say it*

Machine Learning for Text Generation

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

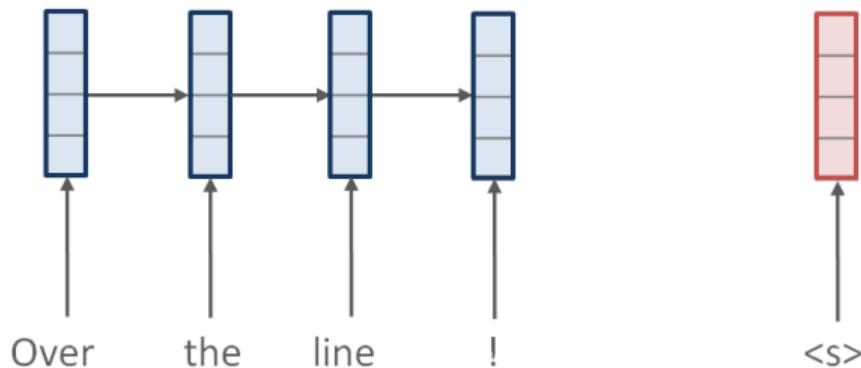
- Input x , *what to talk about*
- Possible output text $y_{1:T}$, *how to say it*
- Scoring function p_θ , with parameters θ learned from data

Machine Translation



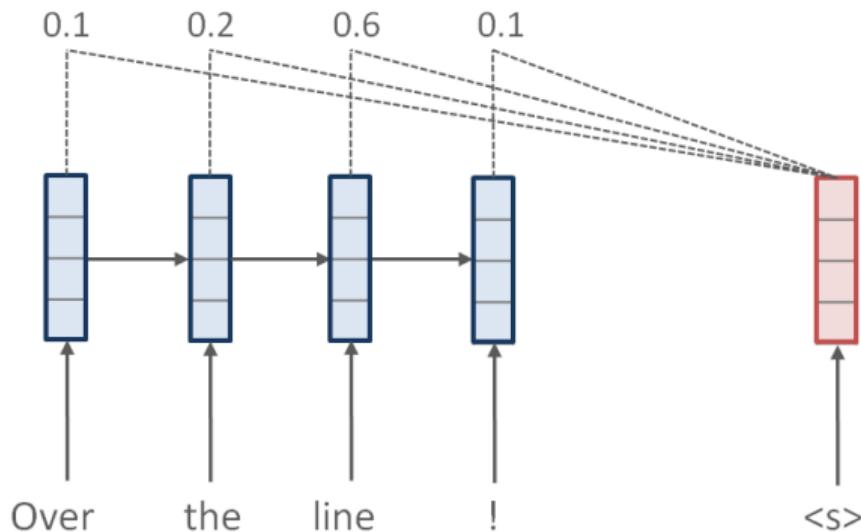
Seq2Seq + Attention

$$p(y_{1:T} \mid x_{1:T}; \theta)$$



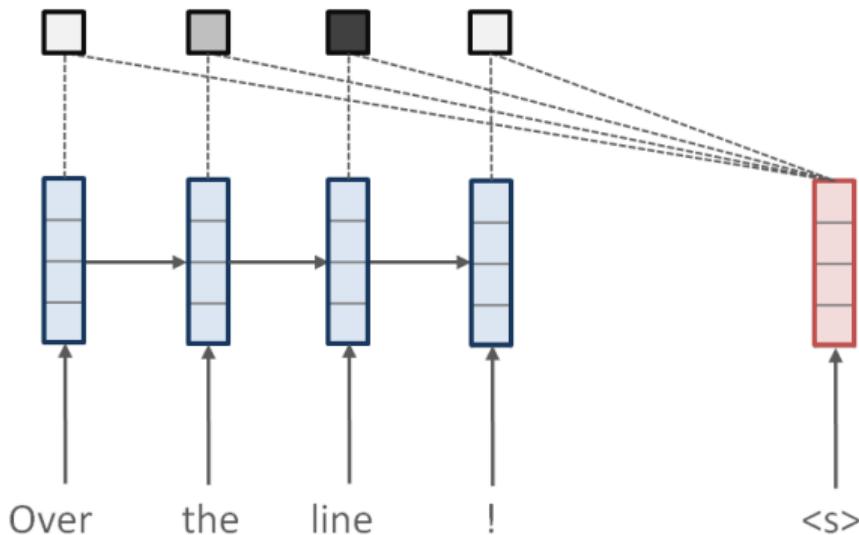
Seq2Seq + Attention

$$p(y_{1:T} \mid x_{1:T}; \theta)$$



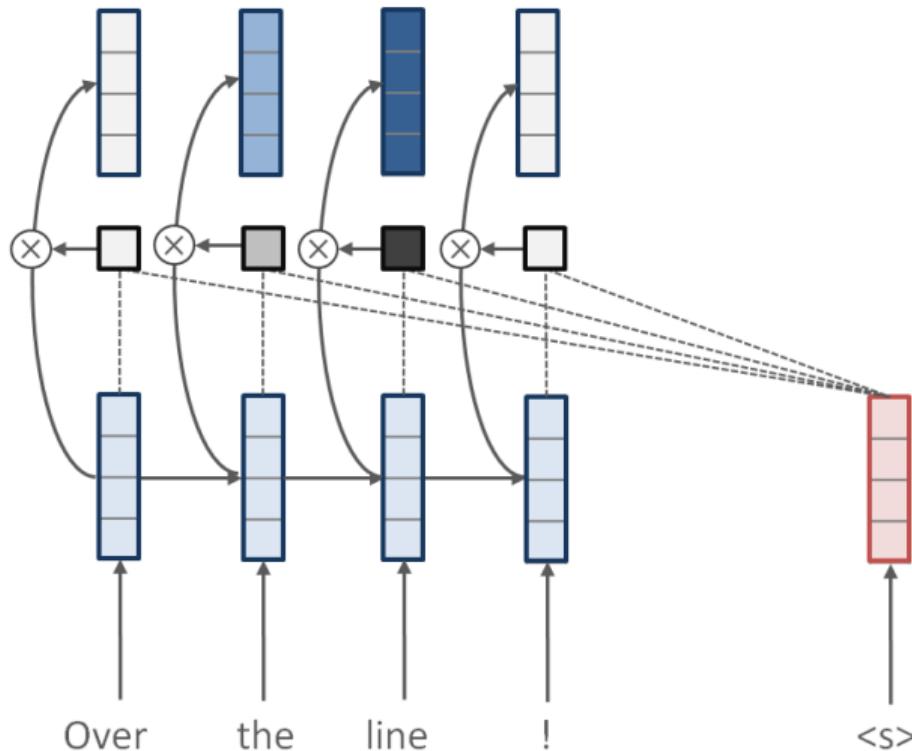
Seq2Seq + Attention

$$p(y_{1:T} \mid x_{1:T}; \theta)$$



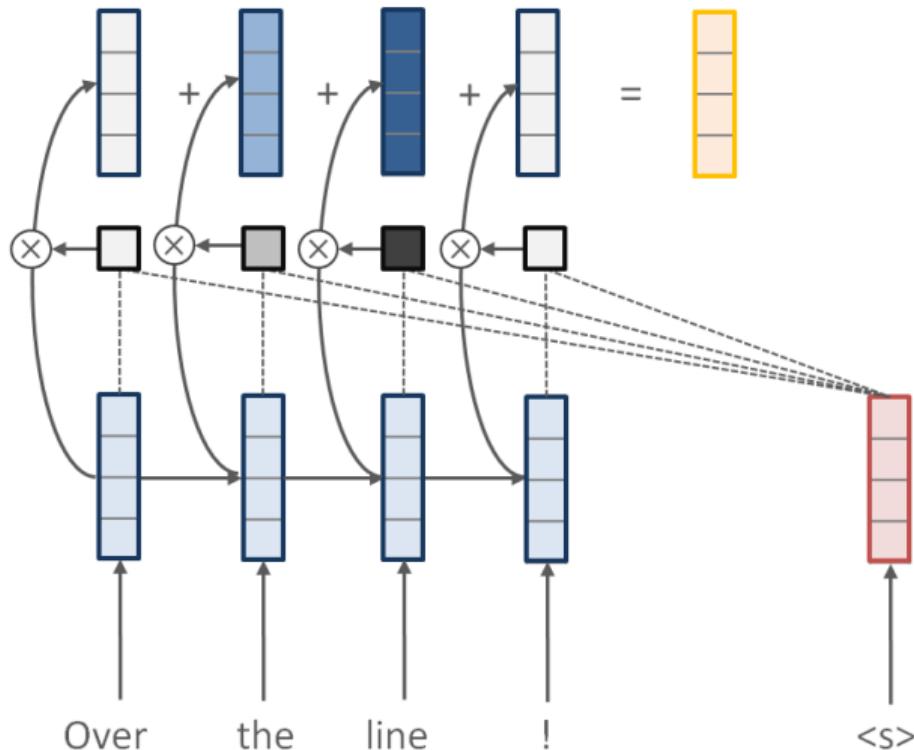
Seq2Seq + Attention

$$p(y_{1:T} \mid x_{1:T}; \theta)$$



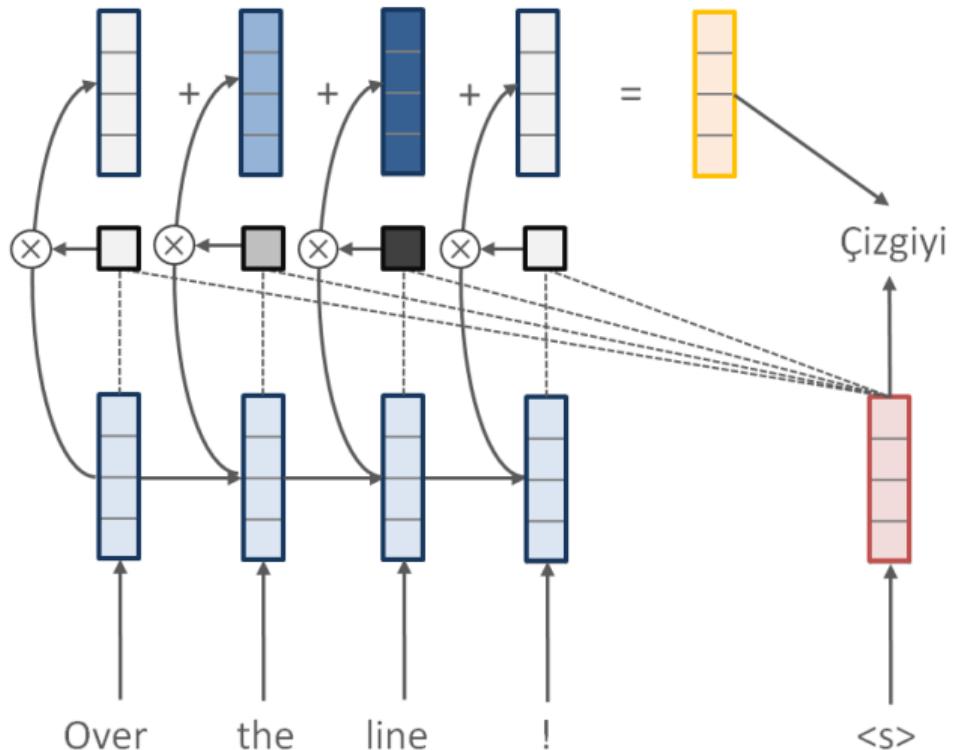
Seq2Seq + Attention

$$p(y_{1:T} \mid x_{1:T}; \theta)$$



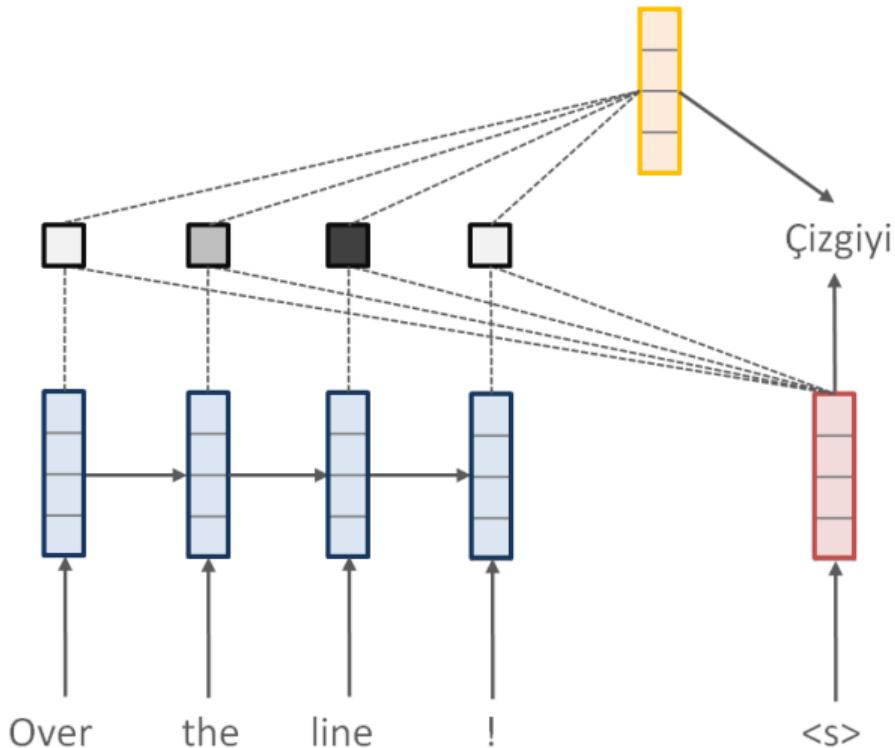
Seq2Seq + Attention

$$p(y_{1:T} \mid x_{1:T}; \theta)$$



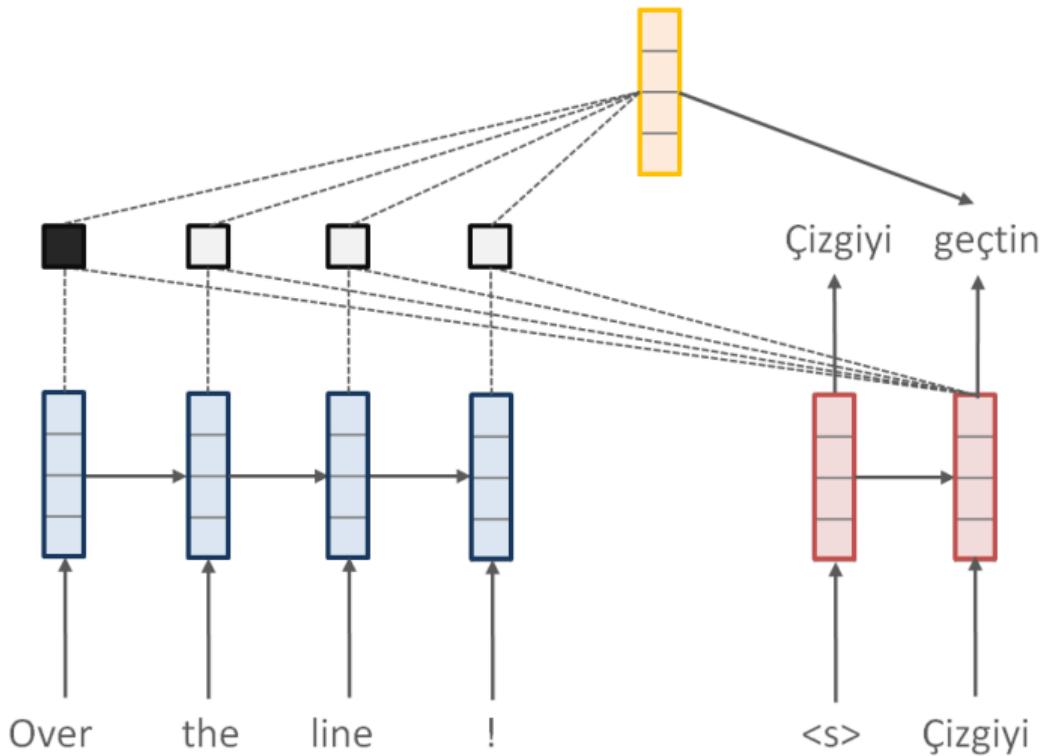
Seq2Seq + Attention

$$p(y_{1:T} \mid x_{1:T}; \theta)$$



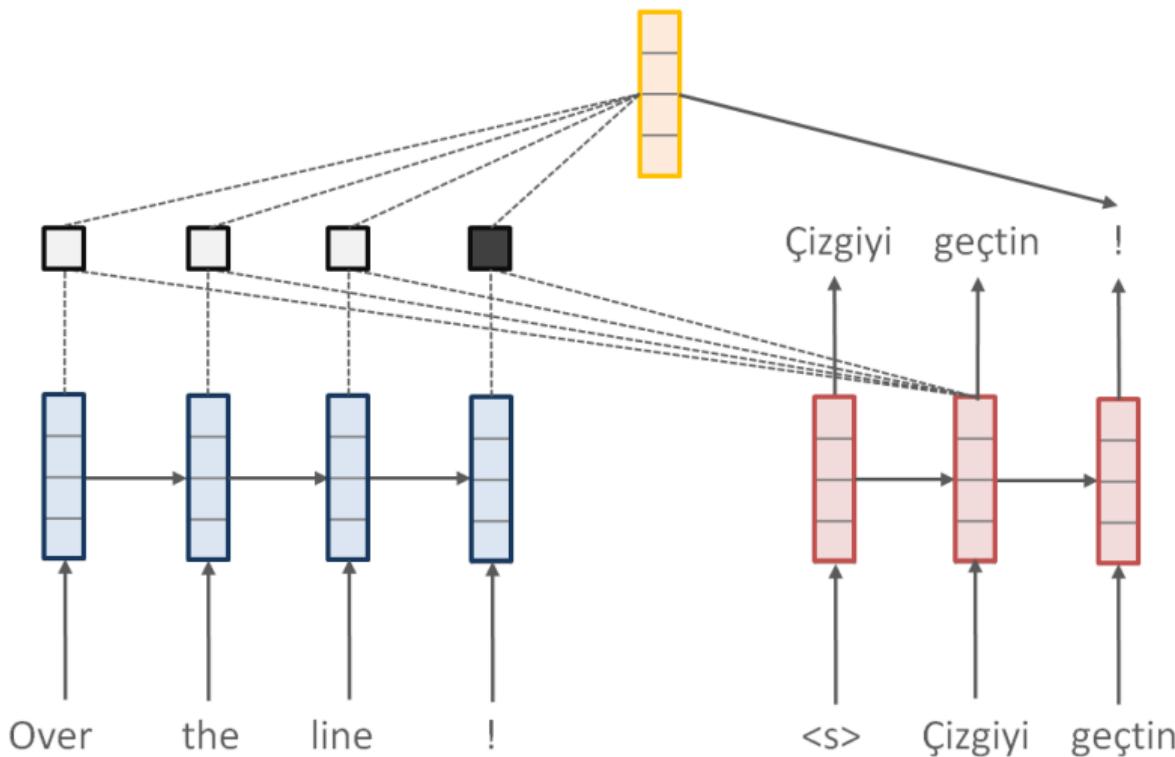
Seq2Seq + Attention

$$p(y_{1:T} \mid x_{1:T}; \theta)$$



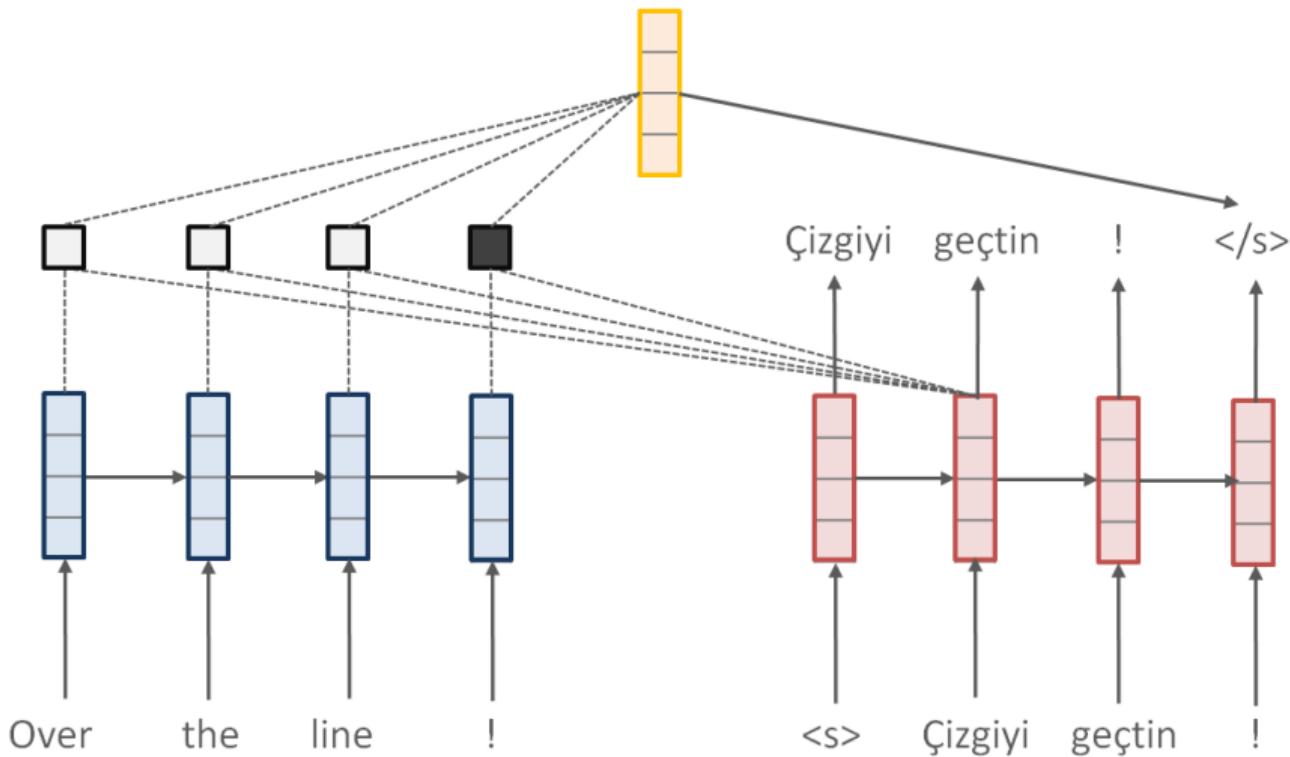
Seq2Seq + Attention

$$p(y_{1:T} \mid x_{1:T}; \theta)$$



Seq2Seq + Attention

$$p(y_{1:T} \mid x_{1:T}; \theta)$$



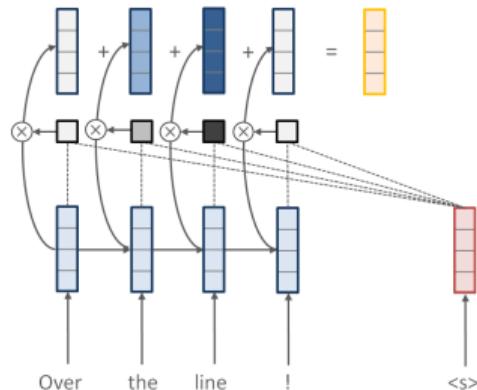
Attention Math

Encoder:

$$\mathbf{h}_s^x \leftarrow \text{RNN}(\mathbf{h}_{s-1}^x, x_s)$$

Attention (Dynamic Context)

$$\alpha \leftarrow \text{softmax}([\mathbf{h}_1^x; \dots; \mathbf{h}_S^x]^\top \mathbf{h}_t) \quad \mathbf{c} \leftarrow \sum_{s=1}^S \alpha_s \mathbf{h}_s^x$$



Attention Math

Encoder:

$$\mathbf{h}_s^x \leftarrow \text{RNN}(\mathbf{h}_{s-1}^x, x_s)$$

Attention (Dynamic Context)

$$\alpha \leftarrow \text{softmax}([\mathbf{h}_1^x; \dots; \mathbf{h}_S^x]^\top \mathbf{h}_t) \quad \mathbf{c} \leftarrow \sum_{s=1}^S \alpha_s \mathbf{h}_s^x$$

Decoder:

$$\mathbf{h}_t \leftarrow \text{RNN}(\mathbf{h}_{t-1}, y_t)$$

Prediction:

$$p(y_{t+1} \mid y_{1:t}, x) = \text{softmax}(\mathbf{W}[\mathbf{h}_t; \mathbf{c}])$$



An open-source neural machine translation system.

English Français 简体中文 한국어
日本語 Русский العربية

Home

[Quickstart \[Lua\]](#)

[Quickstart \[Python\]](#)

[Advanced guide](#)

[Models and Recipes](#)

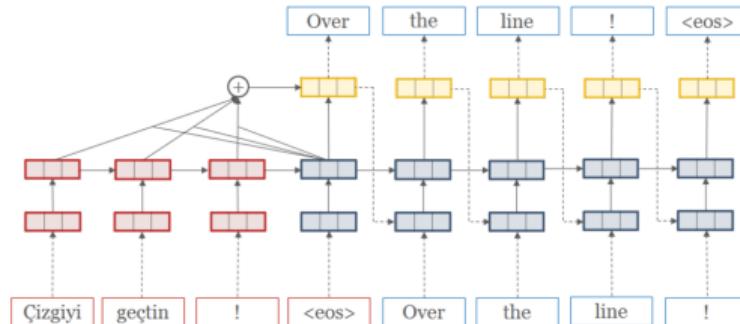
[FAQ](#)

[About](#)

[Documentation](#)

Home

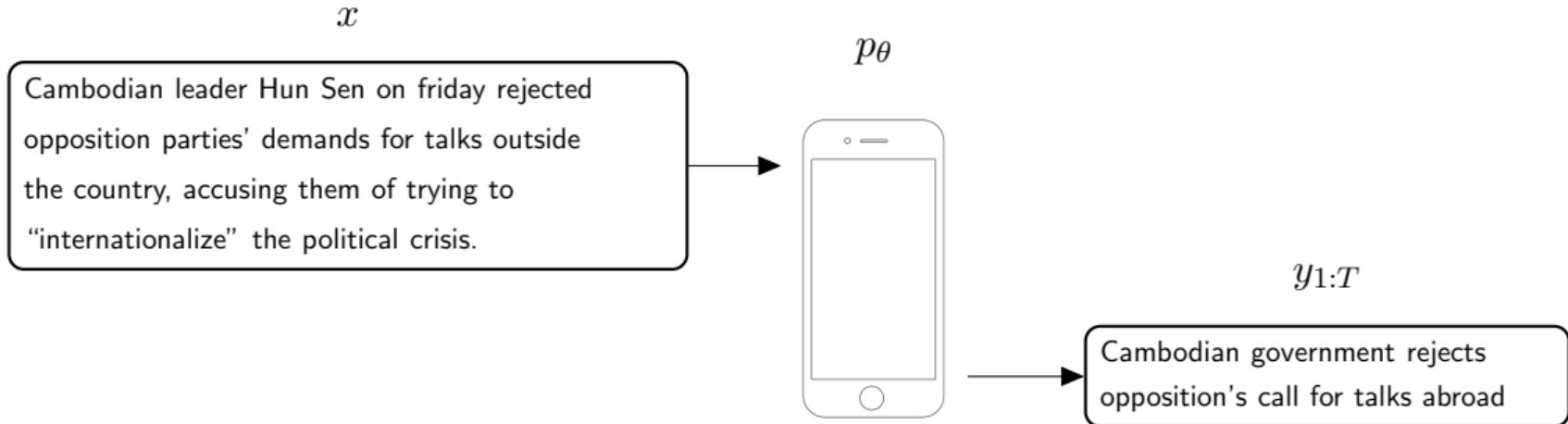
OpenNMT is a industrial-strength, open-source (MIT) neural machine translation system utilizing the [Torch/PyTorch](#) mathematical toolkit.



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.

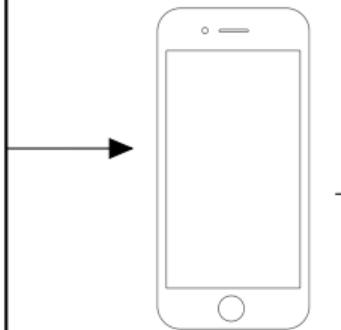
Sentence Summarization

Rush et al. [2015] w/ Facebook



Document Summary

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 the Diagrams

Deng et al. [2016] w/ Bloomberg

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



```
{ \cal K } ^ { L } ( \sigma = 2 ) = \left( \begin{array}{cc} - \frac{d^2}{dx^2} + 4 - \frac{3}{\operatorname{cosh}^2 x} & \frac{3}{\operatorname{cosh}^2 x} \\ \frac{3}{\operatorname{cosh}^2 x} & - \frac{d^2}{dx^2} + 4 - \frac{3}{\operatorname{cosh}^2 x} \end{array} \right)
```

Talk about Data

Wiseman et al. [2017a]

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 ...

What is Missing for Generation?

Building Natural Language Generation Systems (Reiter and Dale. 1999)

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 X parameters these properties will show up.
- **Argument 2:** We need richer models of the underlying phenomenon.
(Also Mirella Lapata, Verena Reiser, Ivan Titov, Shay Cohen and many other folks in the audience!)

Outline

- Background: Core Model and Implementation
- **Work 1: Generation (Learning Neural Templates)**
- Work 2: Attention
- Challenges: Text Generation and Deep Learning

Can we learn to control text generation systems?

Talk about Data

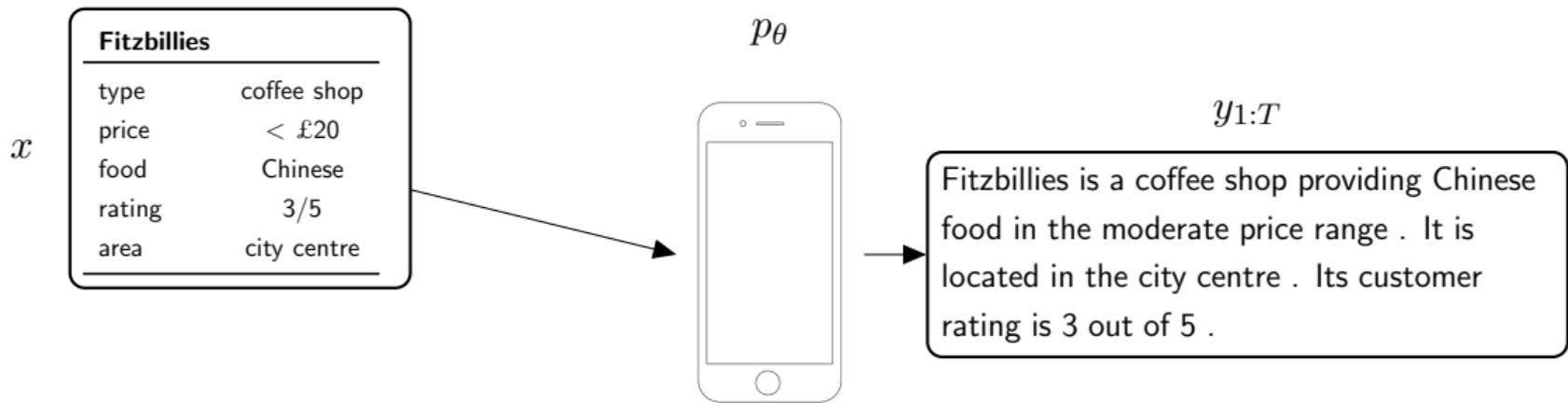
x

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

p_θ



Talk about Data

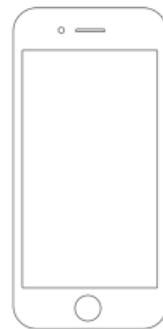


Talking About Data

x



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.	Park.-Rhodes (botany)

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.	Park.-Rhodes (botany)

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

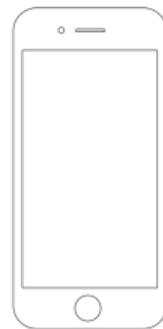
x

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.	Park.-Rhodes (botany)

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.	Park.-Rhodes (botany)

$z_{1:T}$

— (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.

Arguments for Templated Generation

Guarantees about the quality, in particular,

- **Interpretable** in its factual content.
- **Controllable** in terms of style.

Can we achieve these criteria within a deep learning system?

Arguments for Templated Generation

Guarantees about the quality, in particular,

- **Interpretable** in its factual content.
- **Controllable** in terms of style.

Can we achieve these criteria within a deep learning system?

Deep Latent-Variable Models

Strategy: Learn a probabilistic model and *extract* template-like constraints.

Expose specific choices as latent variables z .

$$p(y, z \mid x; \theta)$$

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

Deep Latent-Variable Models

Strategy: Learn a probabilistic model and *extract* template-like constraints.

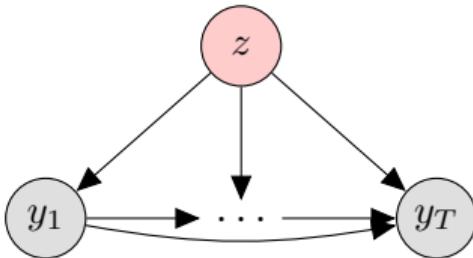
Expose specific choices as latent variables z .

$$p(y, z \mid x; \theta)$$

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

Challenge: Combine with deep learning approach, θ .

Motivating Example: Deep Clustering



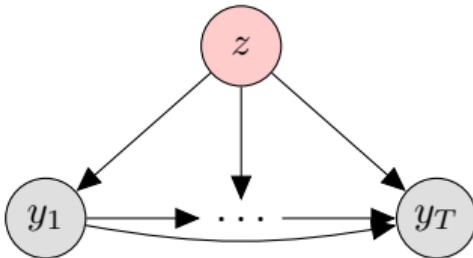
The film is the first from ... $z = 1$

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

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

- ➊ Draw cluster $z \in \{1, \dots, Z\}$.
- ➋ Draw word sequence $y_{1:T}$ from decoder RNN z .

Motivating Example: Deep Clustering



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

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

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

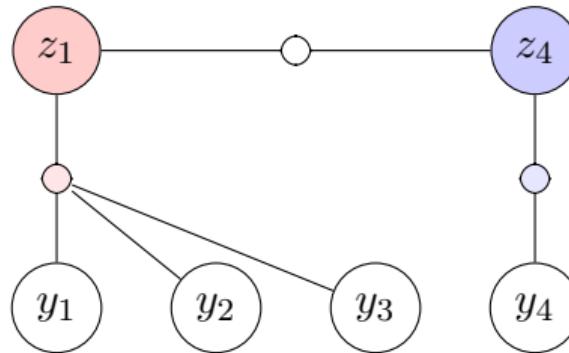
- ➊ Draw cluster $z \in \{1, \dots, Z\}$.
- ➋ Draw word sequence $y_{1:T}$ from decoder RNN z .

Time-Series Clustering

Similar approach can be employed with other probabilistic models.

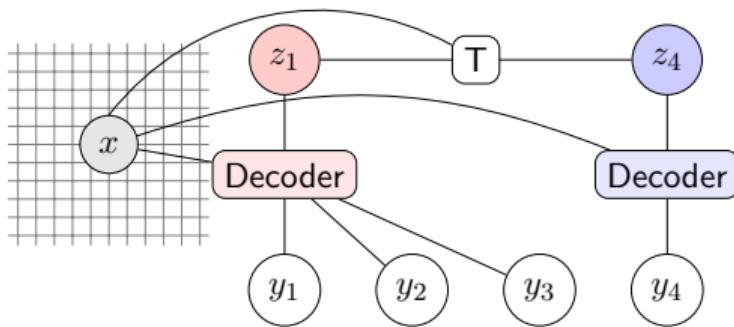
Hidden Semi-Markov Model

- Each discrete cluster produces multiple emissions (e.g. phrases).
- Parameterized with *transition* and *emission* distributions.



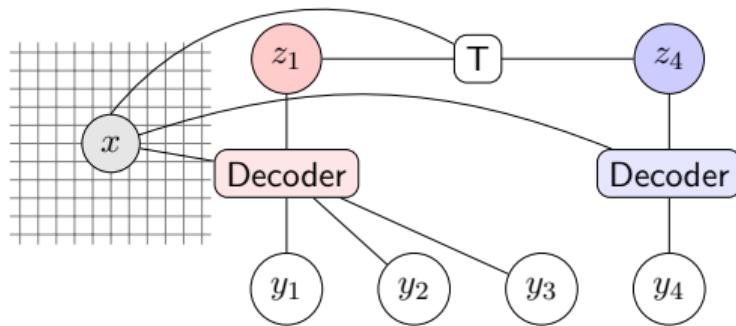
Model: A Deep Hidden Semi-Markov Model

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



Model: A Deep Hidden Semi-Markov Model

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



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.

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

Step 1: Training HSMM

Training requires summing over clusters and segmentation of deep model.

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

Example

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

$$\Downarrow \sum_{z_{1:T}} p(\hat{y}_{1:T}, z_{1:T} | x; \theta)$$

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

Step 1: Training HSMM

Training requires summing over clusters and segmentation of deep model.

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

Example

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

$$\Downarrow \sum_{z_{1:T}} p(\hat{y}_{1:T}, z_{1:T} | x; \theta)$$

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

Step 1: Training HSMM

Training requires summing over clusters and segmentation of deep model.

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

Example

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

$$\Downarrow \sum_{z_{1:T}} p(\hat{y}_{1:T}, z_{1:T} | x; \theta)$$

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 ...

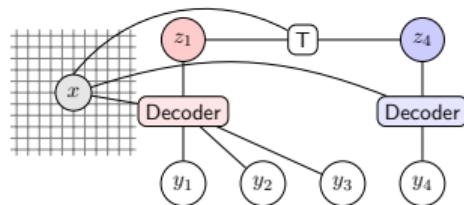
Step 1: Technical Methodology

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.
- Utilize HSMM backward algorithm within standard training.

Step 2: Template Matching

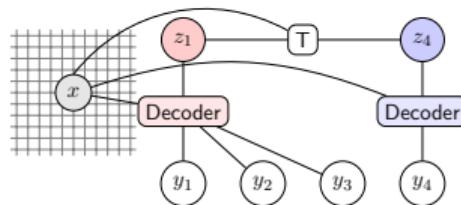
Finding best cluster sequences for each training sentence.



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

Step 2: Template Matching

Finding best cluster sequences for each training sentence.



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

Example

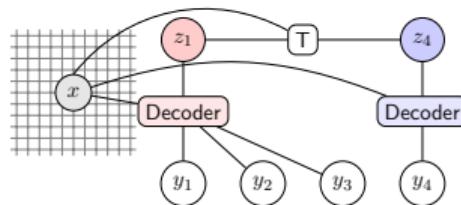
Frederick Parker-Rhodes was an English linguist, plant pathologist

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

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

Step 2: Template Matching

Finding best cluster sequences for each training sentence.



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

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.

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”.

aftab ahmed
anderson da silva | (| born 1951 |) | is an american actor
david jones | ; | born on 1970 | was an american actress |
... | born 1 1974 | is an english cricketer | .

aftab ahmed
anderson da silva | was a world war i member of the austrian house of representatives
david jones | is a former liberal party member of the pennsylvania legislature
... | is a baseball recipient of the montana senate | .

adjutant aftab ahmed
lieutenant anderson da silva | was a world war i member of the knesset
captain david jones | is a former liberal party member of the scottish parliament |
... | is a baseball recipient of the fc lokomotiv liski | .

william " billy " watson 1913 1917
john william smith | (| c. 1900 | in surrey, england | was an american football player
james " jim " edward 1913 | - | british columbia | was an australian rules footballer
... | ... | is an american defenceman | .

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 | (|) | .

Neural Template Generation Approach

x

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

p_{θ}



Neural Template Generation Approach

x

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

$$p_\theta$$



$z_{1:T}$

The _____ is a _____ providing
_____ is an expensive _____ serving
... _____ offering

food high
cuisine moderate
foods less than average

...
price located in the
price range . . It is located near
... near

...
Its customer rating is
Their customer rating is | _____ out of _____
Customers have rated it
...

Neural Template Generation Approach

x

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

$z_{1:T}$

The _____ | is a _____ | providing _____
| is an _____ | is an expensive | _____ | serving _____
| _____ | _____ | _____ | offering _____
| food | _____ | high | _____ | _____ | _____
| cuisine | _____ | in the | moderate | less than average | _____
| foods | _____ | _____ | _____ | _____ | _____
| price | _____ | located in the | _____ | _____ | _____
| price range | . | It is | located near | near | _____ | .
| _____ | _____ | _____ | _____ | _____ | _____
| Its customer rating is | _____ | _____ | _____ | .
| Their customer rating is | _____ | out of | _____ | .
| Customers have rated it | _____ | _____ | .
| _____ | _____ | _____ | _____ | _____ | .

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 || .

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.
-

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

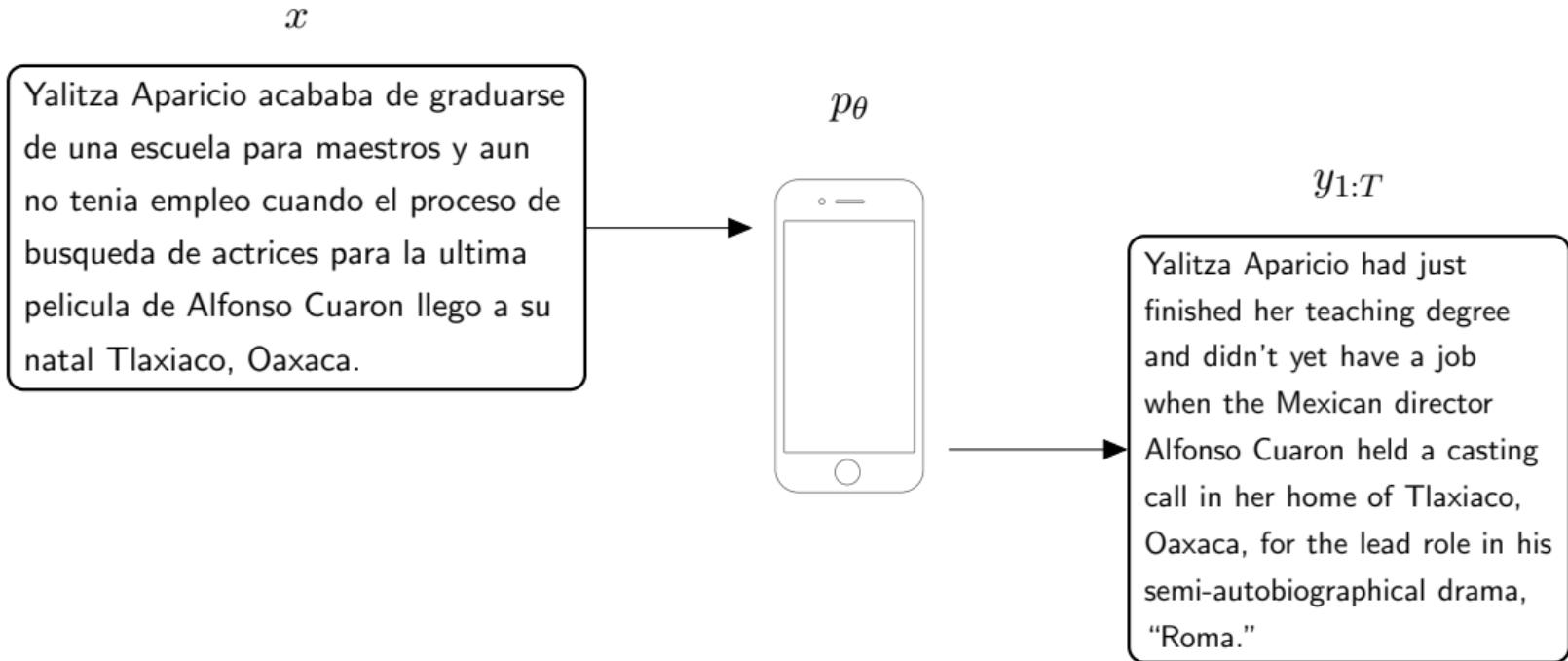
- Encoder/decoder and template-style baselines from Lebret et al. (2016)
- K=45; 1x300 LSTMs as segment/history models
- Used 100 most common $z^{(i)}$ and selected highest overall scorer

Outline

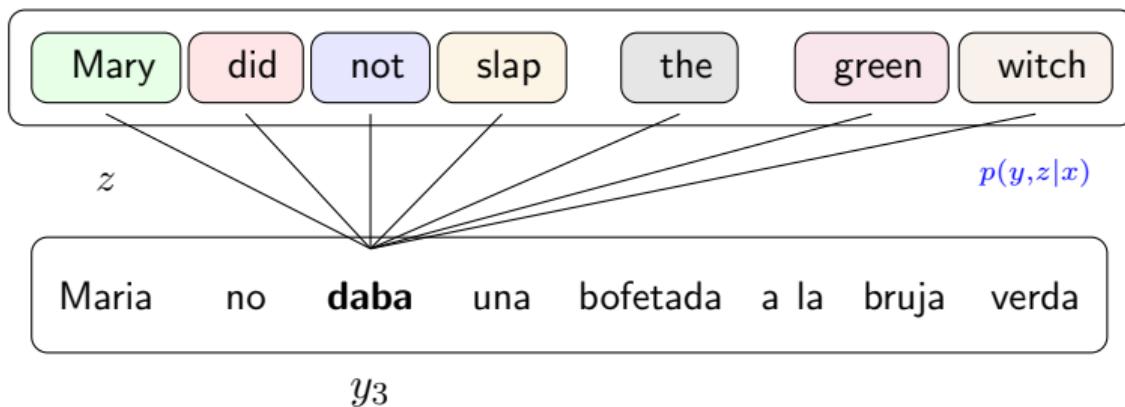
- Background: Core Model and Implementation
- Work 1: Generation
- **Work 2: Attention (Latent Alignment and Variational Attention)**
- Challenges: Text Generation and Deep Learning

Can we learn to control what facts are used?

Machine Learning for Text Generation: Translation



Latent-Variable Alignment Model



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

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
- ...

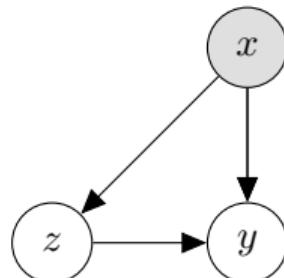
Problem Setup

- Let a be the prior alignment distribution of z
- Let $f(x, z; \theta)$ be the likelihood of y given z

$$z \sim a(x; \theta) \quad y \sim f(x, z; \theta)$$

- Training Objective (maximizing marginal log-likelihood)

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



Key Issue: Computational Cost

- Direct optimization is computationally expensive

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

- Computing expectation requires summing over source for each target.
- Translation bottlenecked by training scale.

Workaround 1: Soft Attention

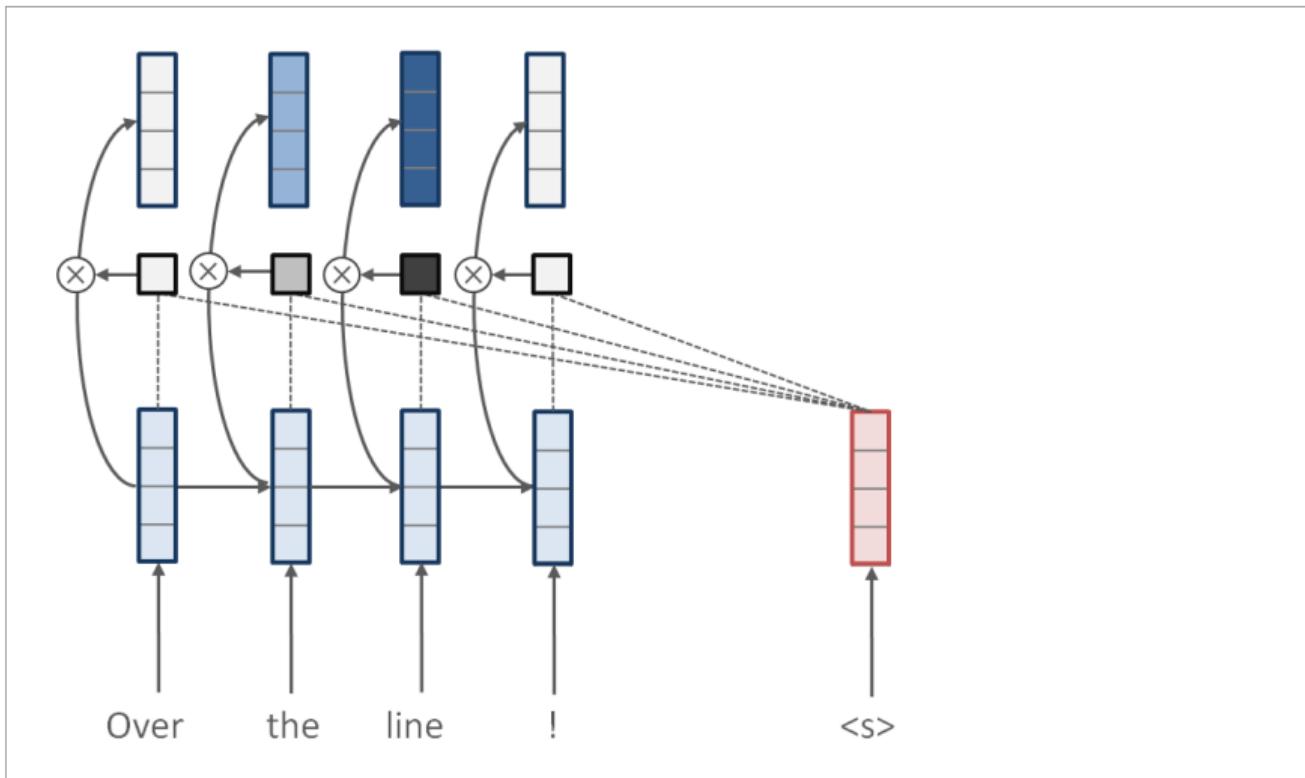
- Replace the joint distribution with a nested expectation [Bahdanau et al 2014]

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

- The corresponding graphical model is



Soft Attention



Workaround 2: Hard Attention

- [Xu et al 2015]: Directly apply Jensen's inequality and optimize with REINFORCE by sampling from the prior

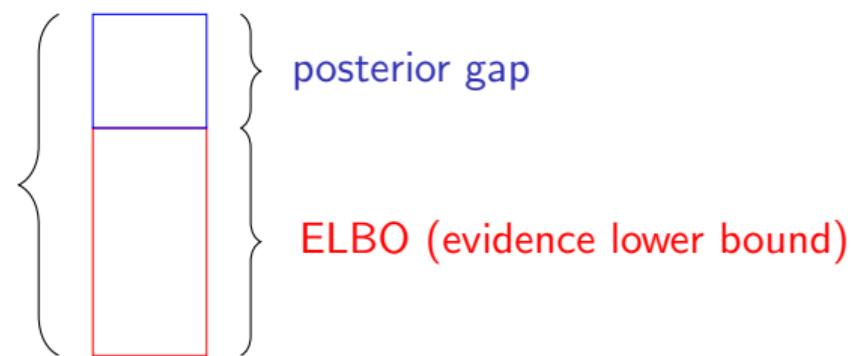
$$\log \mathbb{E}_z[f(x, z)_{\hat{y}}] \geq \mathbb{E}_z \log[f(x, z)_{\hat{y}}] \approx \log f(x, \tilde{z})_{\hat{y}}$$

- Problems:
 - The use of the prior in the expectation may result in a poor bound
 - Cannot directly use for posterior estimation $p(z | y, x)$

Marginal Likelihood: Variational Decomposition

For any* distribution $q(z)$ over z ,

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



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

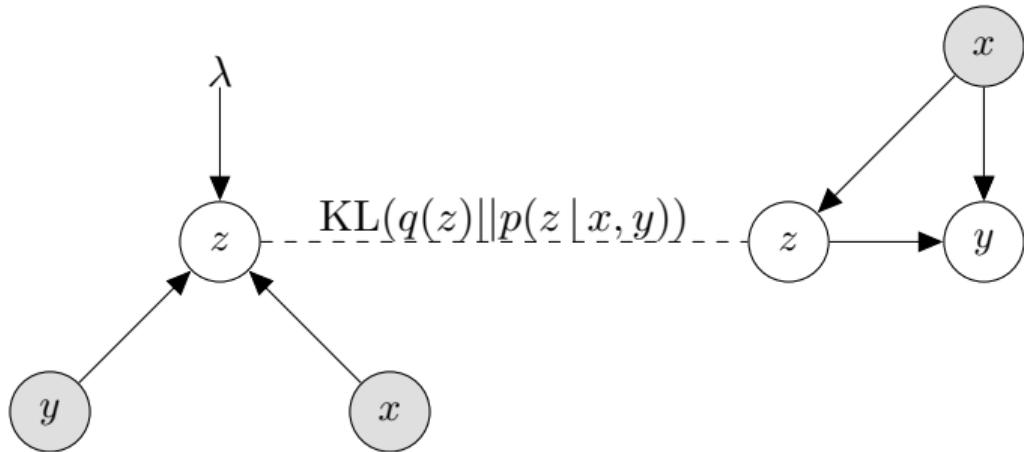
Variational Attention

- Learn model and q to maximize the following lower bound

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

- We choose a $q(z)$ that affords analytic KL
- At test time, marginalize over z .

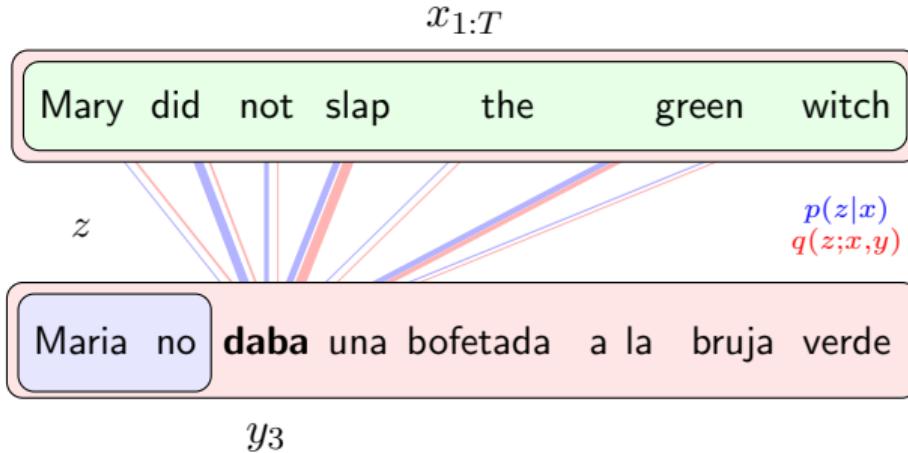
Example Form of q : Amortized Parameterization



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

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

Full Method



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

Technical Details: Categorical and Relaxed

- Categorical (Single Source Alignment Word)

- z and $q(z)$: Categorical Distributions
- Estimate gradients with REINFORCE

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

Technical Details: Categorical and Relaxed

- Categorical (Single Source Alignment Word)

- z and $q(z)$: Categorical Distributions
- 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 and $q(z)$: Dirichlet
- Use reparameterization [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

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

Technical Details: 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

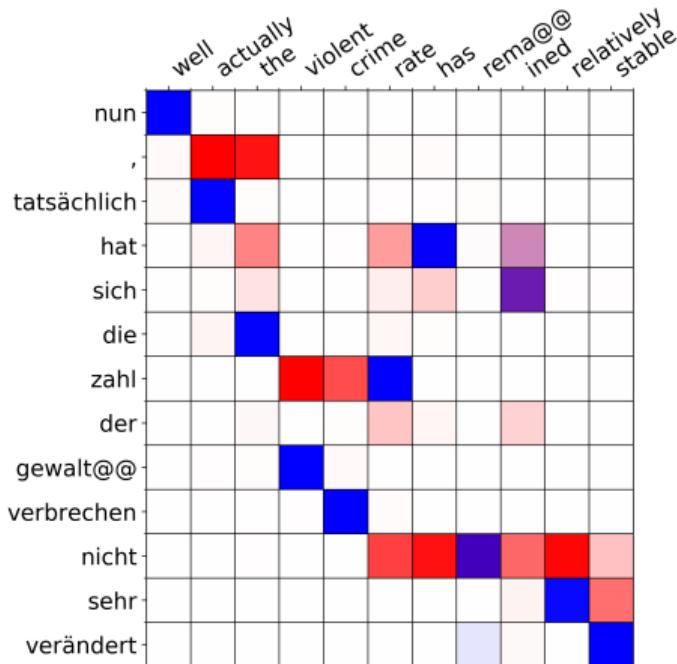
$$\mathbb{E}_{z \sim q(z)}[\nabla_{\theta} \log f(x, z) + (\log f(x, z) - B) \nabla_{\phi} \log q(z)]$$

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

Experiments

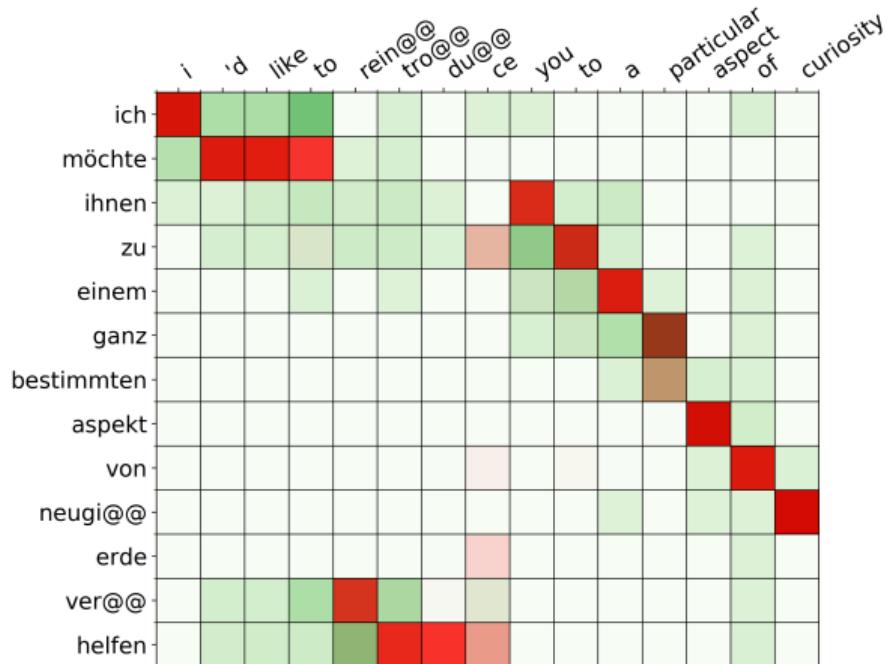
- Full experiments on IWSLT and WMT using LSTM based NMT system.
- Model: Two layer attention based LSTM.
- Variational Model: Bidirectional LSTM model.

Example: Prior / Posterior



Red: prior; blue: posterior.

Example Soft / Latent



Red: prior; green: soft attention.

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

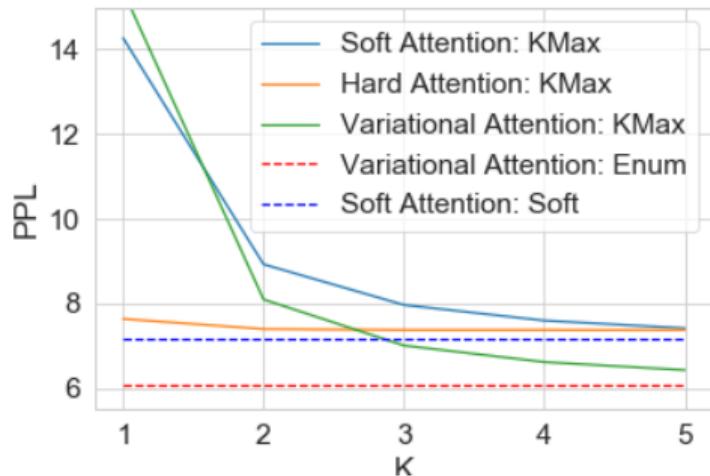
Results (MT: WMT)

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

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

Inference



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

Outline

- Background: Core Model and Implementation
- Work 1: Controlling Generation
- Work 2: Controlling Attention (*Variational Attention*)
- **Challenges: Text Generation and Deep Learning**

Reasoning Systems for Long-Form Generation

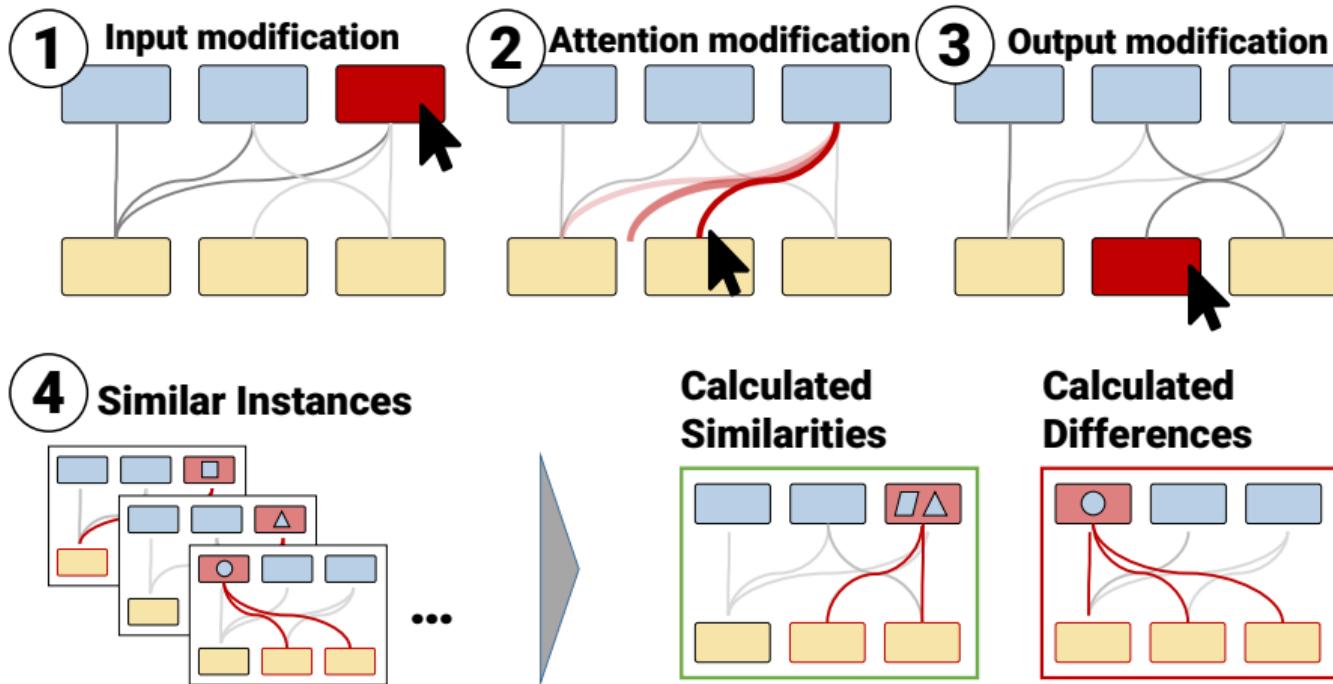
TEAM	WIN	LOSS	PTS	FG_PCT	RB	AS ...
Hawks	11	12	103	49	47	27
Heat	7	15	95	43	34	20
PLAYER	AS	RB	PT	FG	FGA	CITY ...
Tyler Johnson	5	2	27	8	16	Miami
Dwight Howard	11	17	23	9	11	Atlanta
Paul Millsap	2	9	21	8	12	Atlanta
Goran Dragic	4	2	21	8	17	Miami
Wayne Ellington	2	3	19	7	15	Miami
Dennis Schroder	7	4	17	8	15	Atlanta
Rodney McGruder	5	5	11	3	8	Miami
Hasan Whiteside	2	12	8	4	12	Miami
...						

(3) (2) (1)

[The Atlanta Hawks defeated the Miami Heat, 103 - 95, at Philips Arena on Wednesday.] [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.] [Miami (7 - 15) are as beat-up as anyone right now. Hassan Whiteside really struggled in this game, as he amassed eight points, 12 rebounds and one blocks on 4 - of - 12 shooting] ...

Controllable Deep Learning for Translation

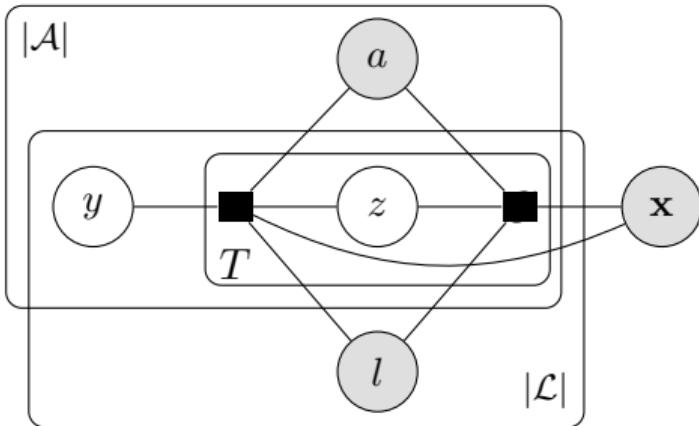
w/ IBM



Prob. Programs for Language Understanding w/ Uber

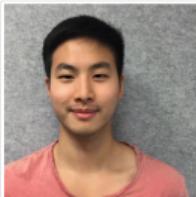


```
def model(z):
    I, J = z.shape
    x = pyro.sample("x", Bernoulli(Px))
    with pyro.plate("I", I, dim=-2):
        y = pyro.sample("y", Bernoulli(Py))
        with pyro.plate("J", J, dim=-1):
            pyro.sample("z", Bernoulli(Pz[x,y]),
                        obs=z)
```



Harvard NLP

Graduate Students



Justin Chiu



Yuntian Deng



Sebastian Gehrmann



Yoon Kim

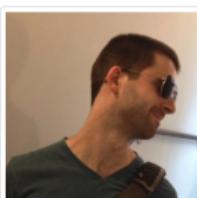


Kelly Zhang



Zachary Ziegler

Grad Alumni



Sam Wiseman
(TTIC)

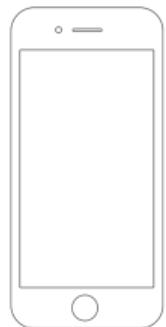
<http://lstm.seas.harvard.edu/client/lstmvis.html?project=00parens&source=states::states2&activation=0.3&cw=30&meta=..&pos=165>

<http://lstm.seas.harvard.edu/client/lstmvis.html?project=05childbook&source=states::states1&activation=0.3&cw=30&meta=..&pos=100&wordBrush=..20,23&wordBrushZero=..1,0&sc=..55,59,159,167,174,179>

Generation with Diagrams

Deng et al. [2016] w/ Bloomberg

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



```
{ \cal K } ^ { L } ( \sigma = 2 ) = \left( \begin{array}{cc} - \frac{d^2}{dx^2} + 4 - \frac{3}{\operatorname{cosh}^2 x} & \frac{3}{\operatorname{cosh}^2 x} \\ \frac{3}{\operatorname{cosh}^2 x} & - \frac{d^2}{dx^2} + 4 - \frac{3}{\operatorname{cosh}^2 x} \end{array} \right)
```

Talk about the Diagrams

Deng et al. [2016] w/ Bloomberg

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



```
{ \cal K } ^ { L } ( \sigma = 2 ) = \left( \begin{array}{cc} - \frac{d^2}{dx^2} + 4 - \frac{3}{\operatorname{cosh}^2 x} & \frac{3}{\operatorname{cosh}^2 x} \\ \frac{3}{\operatorname{cosh}^2 x} & - \frac{d^2}{dx^2} + 4 - \frac{3}{\operatorname{cosh}^2 x} \end{array} \right)
```

Convert images to LaTeX

Take a screenshot of math and paste the LaTeX into your editor, all with a single keyboard shortcut.



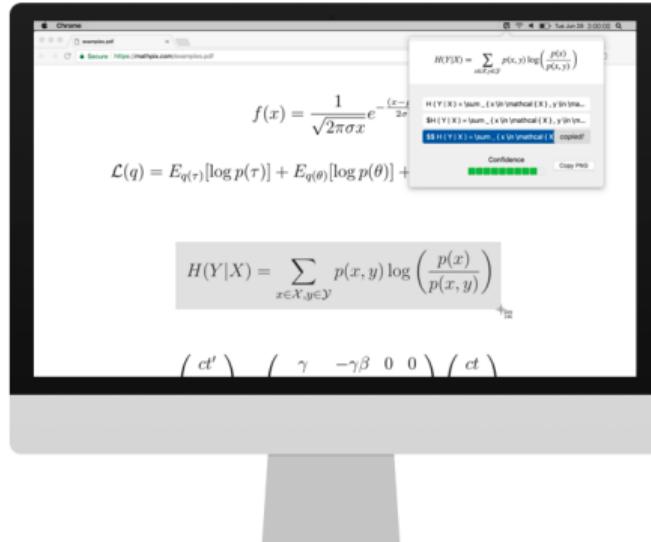
MacOS



Windows



Ubuntu



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 Ramatchandrin,

- Josep Crego, and Alexander Rush. 2018. Opennmt system description for wnmnt 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*.