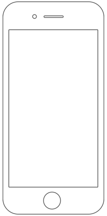
Learning How to Say It: Language Generation and Deep Learning

Alexander M Rush

Machine Learning for Multiclass Classification

x

f_θ



y

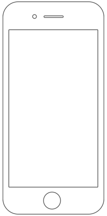
Moped

Machine Learning for Text Generation: Translation

x

Yalitza Aparicio acababa de graduarse de una escuela para maestros y aun no tenia empleo cuando el proceso de busqueda de actrices para la ultima pelicula de Alfonso Cuaron llego a su natal Tlaxiaco, Oaxaca.

f_θ



$y_{1:T}$

Yalitza Aparicio had just finished her teaching degree and didn't yet have a job when the Mexican director Alfonso Cuaron held a casting call in her home of Tlaxiaco, Oaxaca, for the lead role in his semi-autobiographical drama, "Roma."

Machine Learning for Text Generation

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

Machine Learning for Text Generation

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

- Input x, *what to talk about*

Machine Learning for Text Generation

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

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

Machine Learning for Text Generation

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

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

Data-Driven Training and Evaluation

Training :

- Parameters θ learned from a large dataset of paired examples.
- Datasets as large 100k $\rightarrow$ 10 million examples.

Data-Driven Training and Evaluation

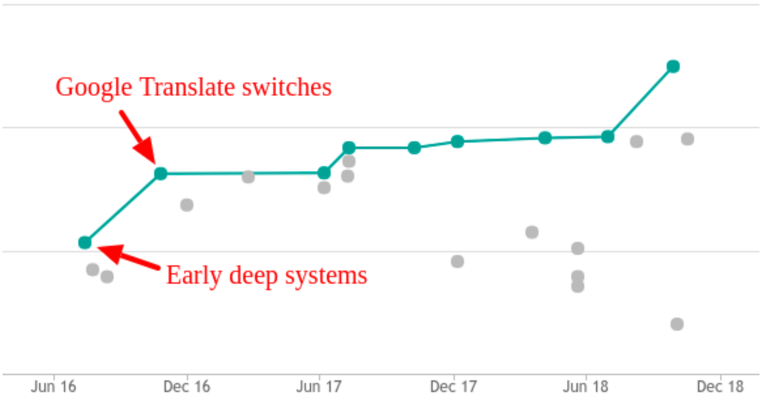
Training :

- Parameters θ learned from a large dataset of paired examples.
- Datasets as large 100k $\rightarrow$ 10 million examples.

Evaluation:

- Truth: `[Yalitza Aparicio had] just [finished her] teaching [degree]`
- Prediction: `[Yalitza Aparicio had] recently [finished her] [degree]`

Translation Performance

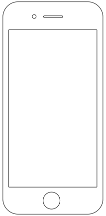


Sentence Summarization

x

Cambodian leader Hun Sen on friday rejected opposition parties' demands for talks outside the country, accusing them of trying to "internationalize" the political crisis.

f_θ

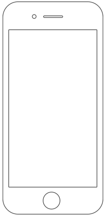


Sentence Summarization

x

Cambodian leader Hun Sen on friday rejected opposition parties' demands for talks outside the country, accusing them of trying to "internationalize" the political crisis.

f_θ



$y_{1:T}$

Cambodian government rejects opposition's call for talks abroad

Summarization

Sep 13, 3:17 PM EDT

GERMANY IMPLEMENTS TEMPORARY BORDER CHECKS TO LIMIT MIGRANTS

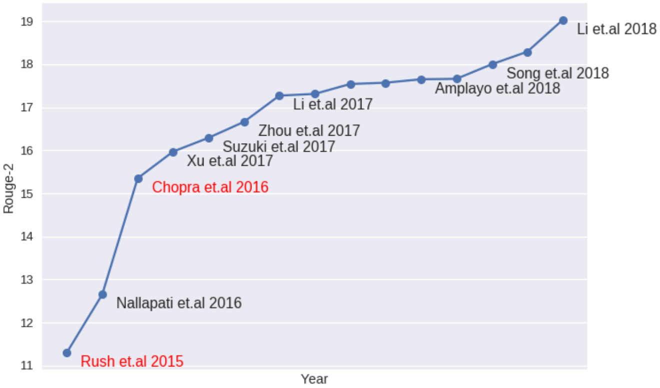
BY GEIR MOULSON AND SHAWN POGATCHNIK
ASSOCIATED PRESS

BERLIN (AP) -- Germany introduced temporary border controls Sunday to stem the tide of thousands of refugees streaming across its frontier, sending a clear message to its European partners that it needs more help with an influx that is straining its ability to cope.

AP Photo/Kay Nietfeld

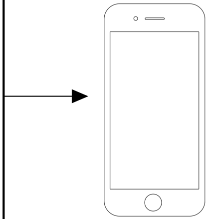
Germany is a preferred destination for many people fleeing Syria's civil war and other troubled nations in the migration crisis that has bitterly divided Europe. They have braved dangerous sea crossings in flimsy

Sentence Summarization Progress



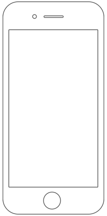
Document Summary

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 . . .



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 ...



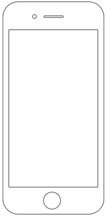
Harry Potter star Daniel Radcliffe gets $20m fortune as he turns 18 monday. Young actor says he has no plans to fritter his cash away. Radcliffe 's earnings from first five potter films have been held in trust fund.

Talk about 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
...						

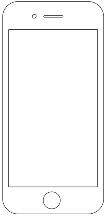


The Atlanta Hawks defeated the Miami Heat, 103 - 95, at Philips Arena on Wednesday. Atlanta was in desperate need of a win and they were able to take care of a short-handed Miami team here. Defense was key for the Hawks, as they held the Heat to 42 percent shooting and forced them to commit 16 turnovers. 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 ...

Talk about the Diagrams

Deng et al. [2016] w/ Bloomberg

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



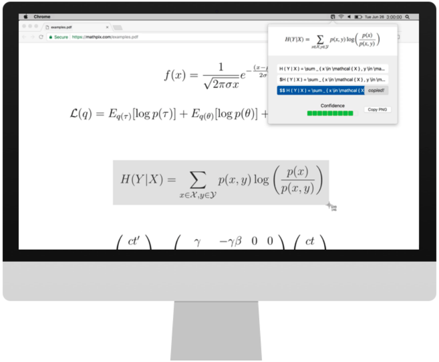
```
{ \cal K } ^ { L } ( \sigma = 2 ) = \left( \begin{array}
{ c c } { - \frac { d ^ { 2 } } { d x ^ { 2 } } + 4 - \frac
{ 3 } { \operatorname { c o s h } ^ { 2 } x } } \& { \frac
{ 3 } { d x ^ { 2 } } }  { \frac { 3 } { \operatorname
{ c o s h } ^ { 2 } x } } \& { - \frac { d ^ { 2 } }
{ d x ^ { 2 }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

Temporary

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



Outline

Goal

Learn How to Say It

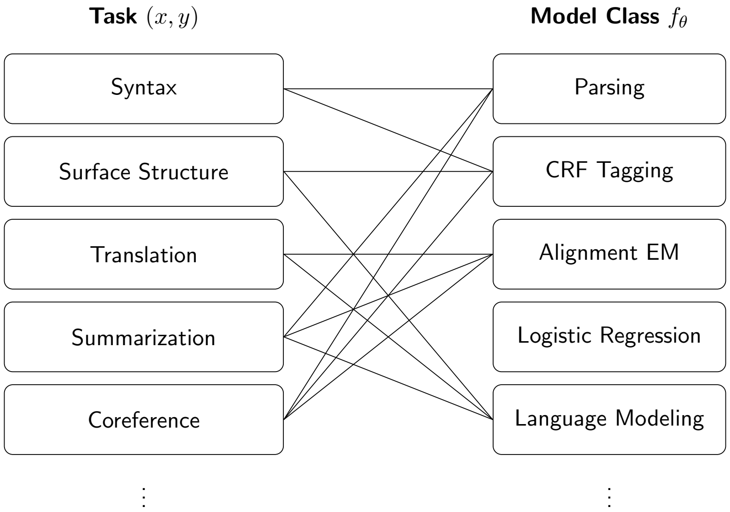
Outline

Goal

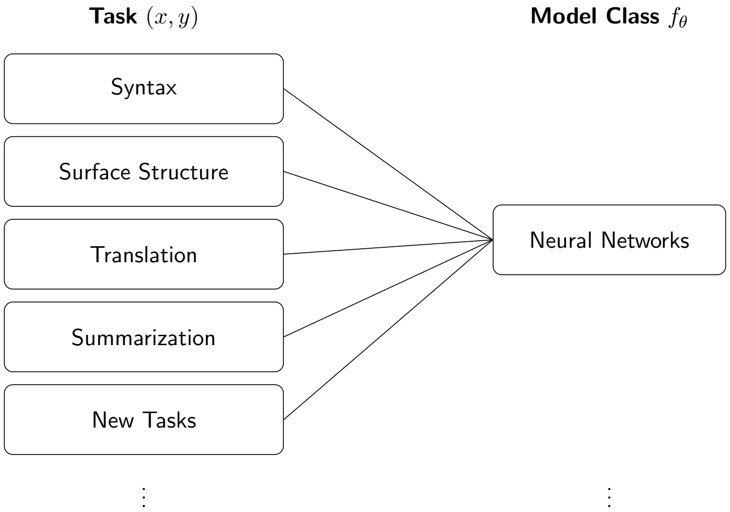
Learn How to Say It

- **Model: Structure and Implementation**
- Work 1: Rethinking Training
- Work 2: Rethinking Generation
- Challenges: Text Generation and Deep Learning

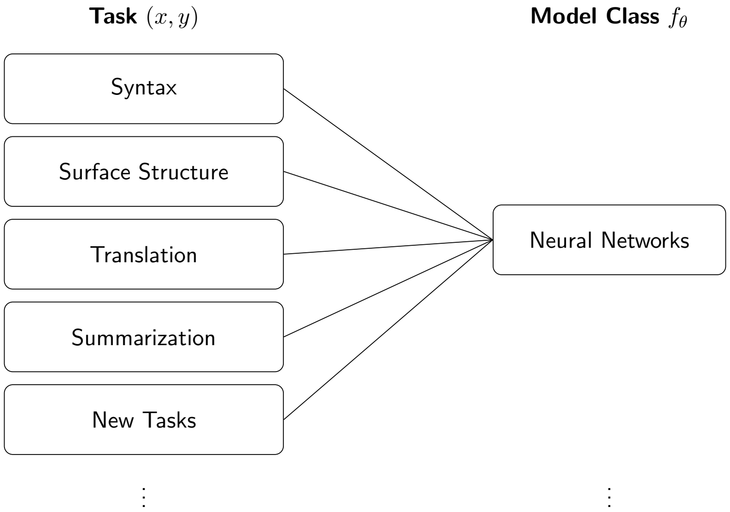
State-of-the-Art Natural Language Processing, circa 2009



State-of-the-Art Natural Language Processing, circa 2019



State-of-the-Art Natural Language Processing, circa 2019



State-of-the-Art Natural Language Processing, circa 2019

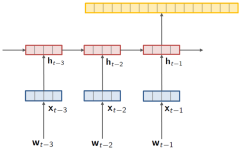
Task (x, y)

- Syntax
- Surface Structure
- Translation
- Summarization
- New Tasks

⋮

Model Class f_θ

Neural Networks



Encoder-Decoder

⋮

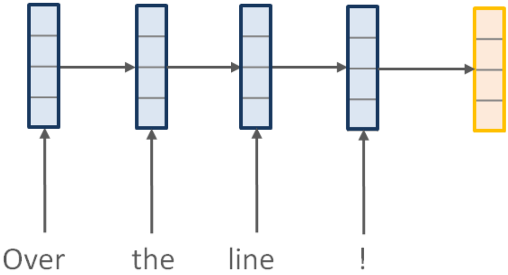
Encoder-Decoder

$f_\theta(y_{1:T}, x_{1:S})$

Over the line !

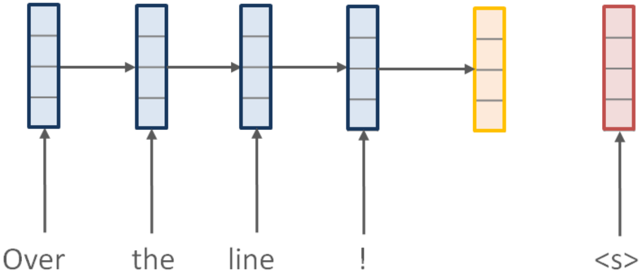
Encoder-Decoder

$f_\theta(y_{1:T}, x_{1:S})$



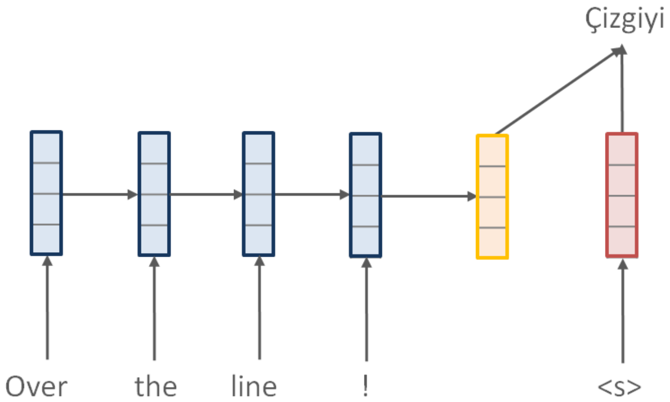
Encoder-Decoder

$f_\theta(y_{1:T}, x_{1:S})$



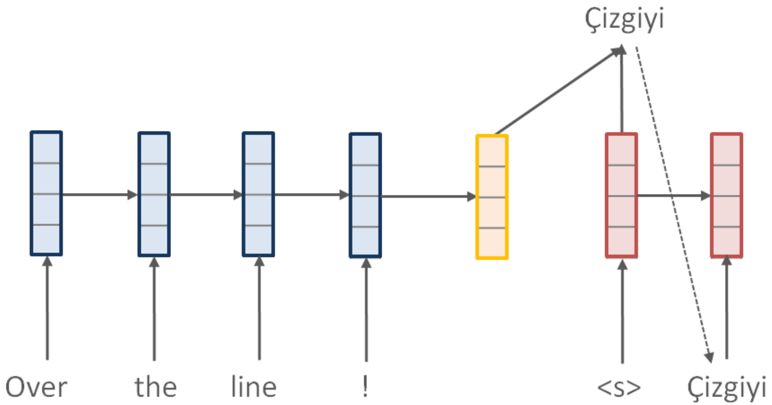
Encoder-Decoder

$$f_\theta(y_{1:T}, x_{1:S})$$



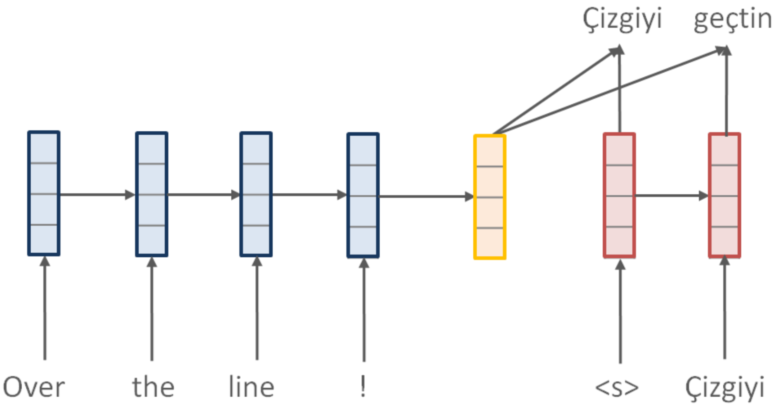
Encoder-Decoder

$$f_\theta(y_{1:T}, x_{1:S})$$



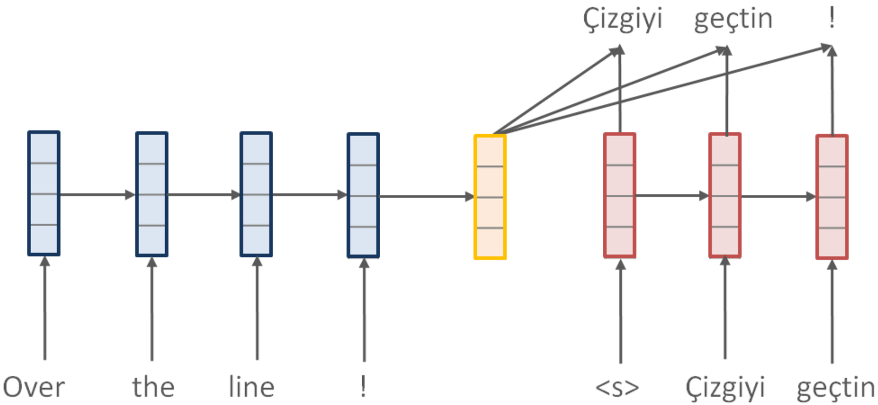
Encoder-Decoder

$f_\theta(y_{1:T}, x_{1:S})$



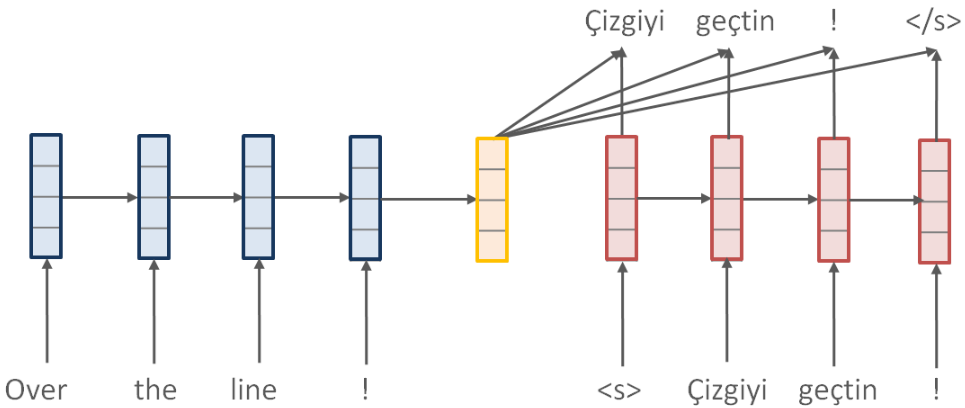
Encoder-Decoder

$f_\theta(y_{1:T}, x_{1:S})$



Encoder-Decoder

$f_\theta(y_{1:T}, x_{1:S})$



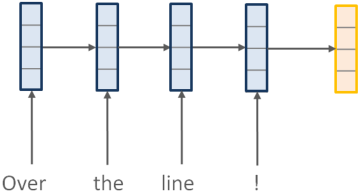
Encoder-Decoder

Encoder:

$$\mathbf{h}^x_s \leftarrow \mathrm{RNN}(\mathbf{h}^x_{s-1}, x_s; \theta)$$

Context:

$$\mathbf{c} = \mathbf{h}^x_S$$



Encoder-Decoder

Encoder:

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

Context:

$$\mathbf{c} = \mathbf{h}_S^x$$

Decoder:

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

Scoring function:

$$p(y_t \mid y_{1:t-1}, x) = \mathrm{softmax}(\mathbf{W}[\mathbf{h}_{t-1}; \mathbf{c}])$$

$$f_\theta(y_{1:T}, x) = \sum_{t=1}^{T} \log p(y_t \mid y_{1:t-1}, x; \theta)$$

Scoring Model Example

Decoder:

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

Language : Well-balanced parentheses (Dyck-1 Language) with nesting-levels,

- Vocabulary: ( ) 0 1 2 3 4
- Example Good String: 0 ((2) (((4 4 4) 3) . . .
- Example Bad String: 0) (3)) ((. . .

LSTMVis - Parenthesis Language Strobelt et al. [2016] w/ IBM

Temporary

LSTMVis - Parenthesis Language Strobelt et al. [2016] w/ IBM

Temporary

Temporary

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

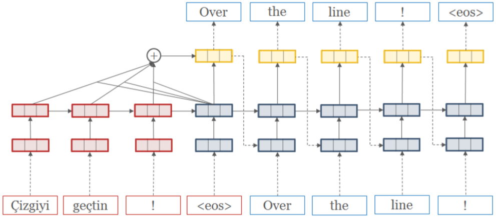
FAQ

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.

Research Overview

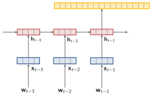
Kim et al. [2016]
Kim et al. [2017]
Wiseman et al. [2017a]

Rush et al. [2015]
Deng et al. [2016]
Schmaltz et al. [2016]

Methods — Applications

Strobelt et al. [2016]
Strobelt et al. [2019]
Wiseman et al. [2017b]

Analysis



Understanding

Wiseman et al. [2018]
Deng et al. [2018]
Kim et al. [2018]

Open-Source — Scaling

Klein et al. [2017]
Senellart et al. [2018]
Rush [2018]

Kim and Rush [2016]
Senellart et al. [2018]
Reagen et al. [2017]

Natural Lang. Processing
Machine Learning
Visualization

Outline

- Model: Structure and Implementation
- **Work 1: Rethinking Training (Beam Search Optimization)**
- Work 2: Rethinking Generation
- Challenges: Text Generation and Deep Learning

Beam Search Optimization

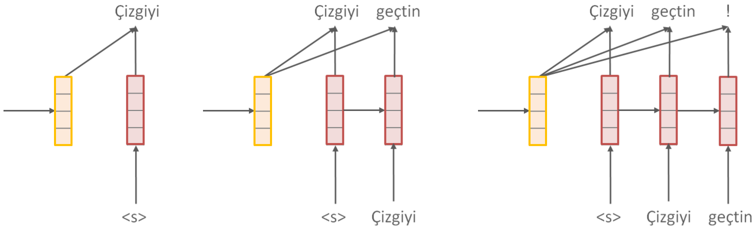
Can we learn parameters θ to better target text generation applications?

Training Setup:

- $(x, \hat{y}_{1:T})$ - input, output sentence pair
- $\mathcal{L}(\theta)$ - loss function
- f_θ - learned scoring function

Baseline: Training Encoder-Decoder

Parameters θ are trained to score the next word given the *true* history, $\hat{y}_{1:t-1}$

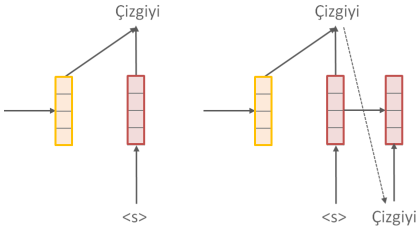


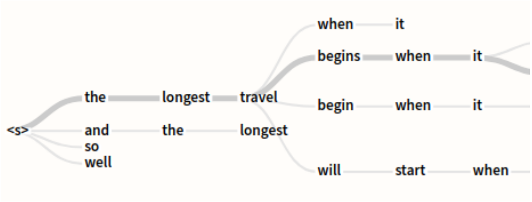
Training loss is identical to multiclass classification,

$$\mathcal{L}(\theta) = -\sum_{t} \log p(\hat{y}_t \mid \hat{y}_{1:t-1}, x; \theta)$$

Generating with Encoder-Decoder

Parameters θ are used to score the next word given *any* history, $y_{1:t-1}$.





Generation aims to maximize over all sequences,

$$y^*_{1:T} = \arg\max_{y_{1:T}} f_\theta(y_{1:T}, x) = \arg\max_{y_{1:T}} \sum_t \log p(y_t | y_{1:t-1}, x; \theta)$$

Standard Heuristic Method: Beam Search

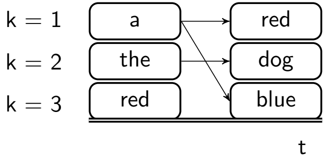
$$y^*_{1:T} \approx \arg\max_{y_{1:T}} f_\theta(y_{1:T}, x)$$

k = 1	a
k = 2	the
k = 3	red

t

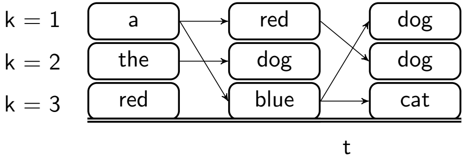
Standard Heuristic Method: Beam Search

$$y^*_{1:T} \approx \arg\max_{y_{1:T}} f_\theta(y_{1:T}, x)$$



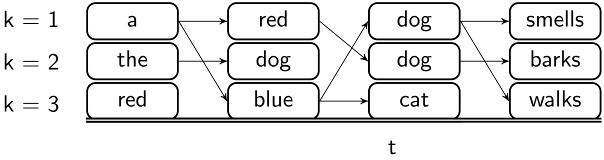
Standard Heuristic Method: Beam Search

$$y^*_{1:T} \approx \arg\max_{y_{1:T}} f_\theta(y_{1:T}, x)$$



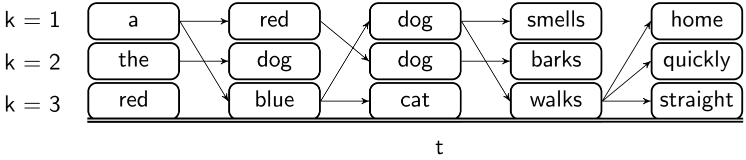
Standard Heuristic Method: Beam Search

$$y^*_{1:T} \approx \arg\max_{y_{1:T}} f_\theta(y_{1:T}, x)$$



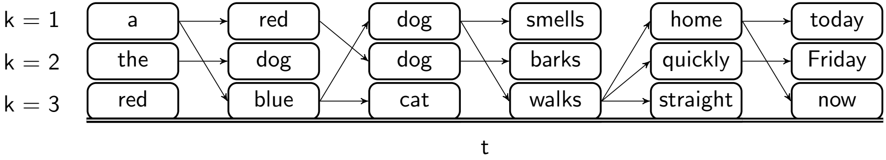
Standard Heuristic Method: Beam Search

$$y^*_{1:T} \approx \arg\max_{y_{1:T}} f_\theta(y_{1:T}, x)$$



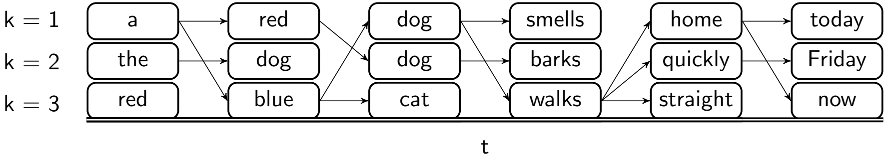
Standard Heuristic Method: Beam Search

$$y^*_{1:T} \approx \arg\max_{y_{1:T}} f_\theta(y_{1:T}, x)$$



Standard Heuristic Method: Beam Search

$$y^*_{1:T} \approx \arg\max_{y_{1:T}} f_\theta(y_{1:T}, x)$$



1. Compute the score of every hypothesis k and possible next word y_t,

$$f_\theta(\langle y_t, y^{(k)}_{1:t-1}\rangle, x) \leftarrow \log p(y_t \mid y^{(k)}_{1:t-1}, x) + \log p(y^{(k)}_{1:t-1} \mid x)$$

2. Prune to only the K highest-scoring of $(K \times \text{vocab})$ choices,

$$y^{(1:K)}_{1:t} \leftarrow K \operatorname*{arg\,max}_{y_t, k} f_\theta(\langle y_t, y^{(k)}_{1:t}\rangle, x)$$

Core Issues

Multiclass Training $\Rightarrow$ Structured Generation ?

Core Issues

Multiclass Training $\Rightarrow$ Structured Generation ?

1. Exposure Bias
   - Training conditions on true history, but generation uses predicted history.

Core Issues

Multiclass Training $\Rightarrow$ Structured Generation ?

1. Exposure Bias
   - Training conditions on true history, but generation uses predicted history.
2. Label Bias
   - Training is locally multiclass, but score is over entire sequences.

Core Issues

Multiclass Training $\Rightarrow$ Structured Generation ?

1. Exposure Bias
   - Training conditions on true history, but generation uses predicted history.
2. Label Bias
   - Training is locally multiclass, but score is over entire sequences.
3. Metric Bias
   - Training uses multiclass classification, but evaluation uses n-gram match.

Core Issues

Multiclass Training $\Rightarrow$ Structured Generation ?

1. Exposure Bias
   - Training conditions on true history, but generation uses predicted history.
2. Label Bias
   - Training is locally multiclass, but score is over entire sequences.
3. Metric Bias
   - Training uses multiclass classification, but evaluation uses n-gram match.

Strategy: Modify training to fix these issues.

Modification 1: Beam Search at Training

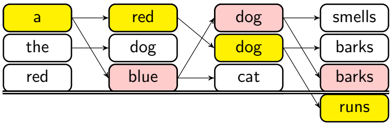
Fix: Exposure Bias

- Take prediction algorithm into account during training.

Modification 1: Beam Search at Training

Fix: Exposure Bias

- Take prediction algorithm into account during training.



- Run our beam search procedure during training (structured training)

- Loss tied to mistakes, e.g. true sequence $\hat{y}_{1:t}$ is *violated* by $y_{1:t}^{(K)}$ worst beam

Modification 2: Global Scoring Function

Fix: Label Bias

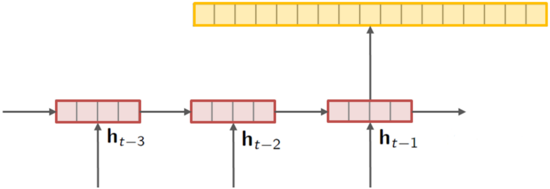
- Use a global sequence scoring function.

Modification 2: Global Scoring Function

Fix: Label Bias

- Use a global sequence scoring function.

$$f_\theta(y_{1:t}, x) = \mathbf{W}[\mathbf{h}_{t-1}; \mathbf{c}]$$



- Replace local $\log p(y_t | y_{1:t-1}, x; \theta)$ with a global scoring model $f_\theta(y_{1:t}, x)$.

Modification 3: Train with Margin

Fix: Metric Bias

- Incorporate a metric specific term, e.g. n-gram mismatch

Modification 3: Train with Margin

Fix: Metric Bias

- Incorporate a metric specific term, e.g. n-gram mismatch

$$\mathcal{L}(\theta) = \sum_t \Delta(\hat{y}_{1:t}, y_{1:t}^{(K)}) \left[1 - f_\theta(\hat{y}_{1:t}, x) + f_\theta(y_{1:t}^{(K)}, x) \right]$$

- Positive if true sequence $\hat{y}_{1:t}$ within margin of worst beam sequence $y_{1:t}^{(K)}$
- Slack-rescaled margin takes problem-specific Δ into account

Extension: Training with Hard Constraints

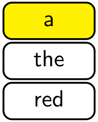
Bonus: Hard Constraints

- Beam Search Optimization allows users to enforce hard constraints at training.

Example: Code generation with a known grammar,

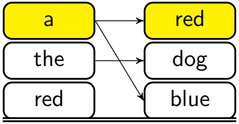
```
{ \cal K } ^ { L } ( \sigma = 2 ) = \left( \begin{array}
{ c c } { - \frac { d ^ { 2 } } { d x ^ { 2 } } + 4 - \frac
{ 3 } { \operatorname { c o s h } ^ { 2 } x } } \& { \frac
{ 3 } { d x ^ { 2 } } }  { \frac { 3 } { \operatorname
{ c o s h } ^ { 2 } x } } \& { - \frac { d ^ { 2 } }
{ d x ^ { 2 }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

Beam Search Optimization: Training Example



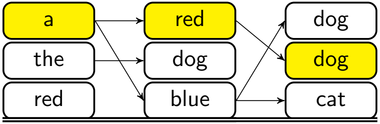
- True: ground-truth sequence $\hat{y}_{1:t}$

Beam Search Optimization: Training Example



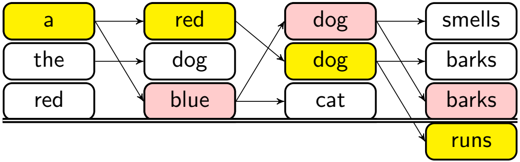
- True: ground-truth sequence $\hat{y}_{1:t}$

Beam Search Optimization: Training Example



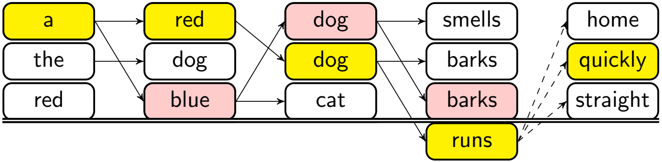
- True: ground-truth sequence $\hat{y}_{1:t}$

Beam Search Optimization: Training Example



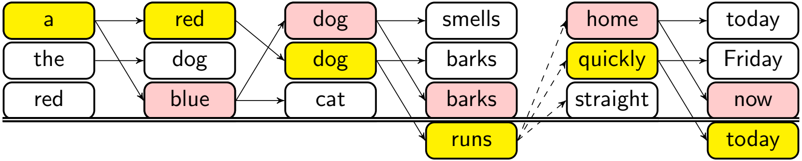
- True: ground-truth sequence $\hat{y}_{1:t}$
- Predicted: lowest-scoring violating sequence $y_{1:t}^{(K)}$

Beam Search Optimization: Training Example



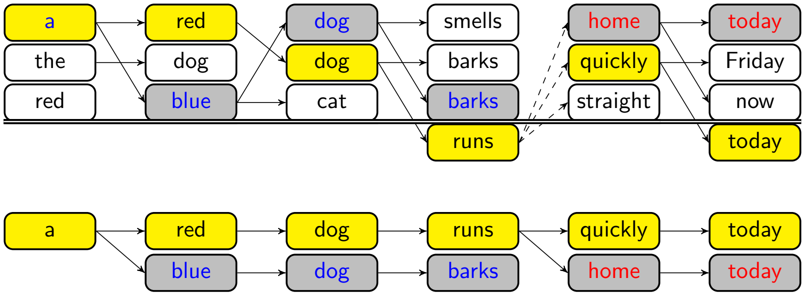
- True: ground-truth sequence $\hat{y}_{1:t}$
- Predicted: lowest-scoring violating sequence $y_{1:t}^{(K)}$

Beam Search Optimization: Training Example



- True: ground-truth sequence $\hat{y}_{1:t}$
- Predicted: lowest-scoring violating sequence $y_{1:t}^{(K)}$

Parameter Updates: Structured Backpropagation



Main Results

Train Beam	$K = 1$	$K = 5$	$K = 10$
	Word Ordering (BLEU)		
Encoder-Decoder	25.2	29.8	31.0
Beam Search Optimization	28.0	33.2	34.3
Beam Search Optimization-Constraints	**28.6**	**34.3**	**34.5**

Main Results

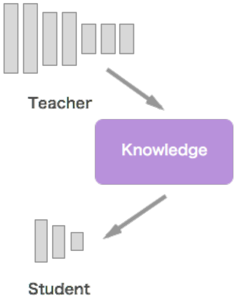
Train Beam	$K = 1$	$K = 5$	$K = 10$
	Word Ordering (BLEU)		
Encoder-Decoder	25.2	29.8	31.0
Beam Search Optimization	28.0	33.2	34.3
Beam Search Optimization-Constraints	**28.6**	**34.3**	**34.5**
	Dependency Parsing (UAS)		
Encoder-Decode	**87.33**	88.53	88.66
Beam Search Optimization	86.91	91.00	91.17
Beam Search Optimization-Constraints	85.11	**91.25**	**91.57**

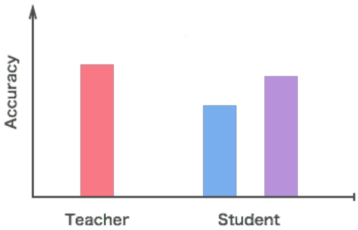
Main Results

Train Beam	$K = 1$	$K = 5$	$K = 10$
	Word Ordering (BLEU)		
Encoder-Decoder	25.2	29.8	31.0
Beam Search Optimization	28.0	33.2	34.3
Beam Search Optimization-Constraints	**28.6**	**34.3**	**34.5**
	Dependency Parsing (UAS)		
Encoder-Decode	**87.33**	88.53	88.66
Beam Search Optimization	86.91	91.00	91.17
Beam Search Optimization-Constraints	85.11	**91.25**	**91.57**
	Machine Translation (BLEU)		
Encoder-Decoder	22.53	24.03	23.87
Beam-Search Optimization, Δ	**23.83**	**26.36**	**25.48**

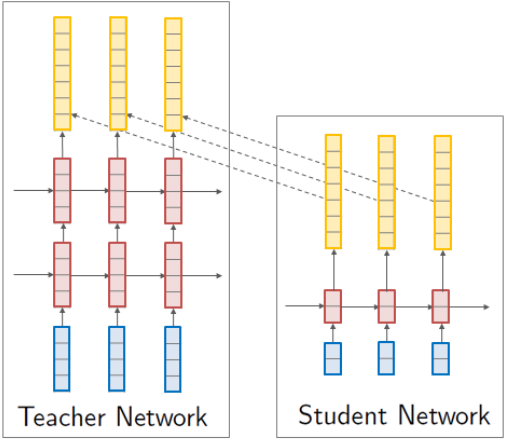
Goal: Shrink the size of text generation models.

- Knowledge Distillation: Train a *student* model to learn from a *teacher* model.

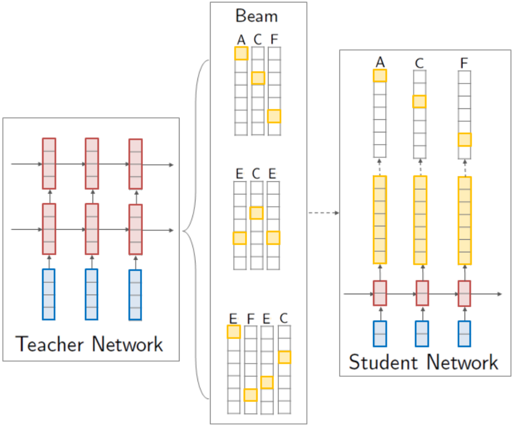




Word-Level Knowledge Distillation (Multiclass)



Sequence-Level Knowledge Distillation



Results: WMT English $\rightarrow$ German Translation

Model	BLEU$_{K=1}$	$\Delta_{K=1}$	BLEU$_{K=5}$	$\Delta_{K=5}$
4×1000				
Teacher	17.7	$-$	19.5	$-$

Results: WMT English $\rightarrow$ German Translation

Model	BLEU$_{K=1}$	$\Delta_{K=1}$	BLEU$_{K=5}$	$\Delta_{K=5}$
4×1000				
Teacher	17.7	$-$	19.5	$-$
2×500				
Student	14.7	$-$	17.6	$-$
Word-KD	15.4	$+0.7$	17.7	$+0.1$

Results: WMT English $\rightarrow$ German Translation

Model	BLEU$_{K=1}$	$\Delta_{K=1}$	BLEU$_{K=5}$	$\Delta_{K=5}$
4×1000				
Teacher	17.7	$-$	19.5	$-$
2×500				
Student	14.7	$-$	17.6	$-$
Word-KD	15.4	$+0.7$	17.7	$+0.1$
Seq-KD	18.9	$+\mathbf{4.2}$	19.3	$+\mathbf{1.7}$

Application

Temporary

Outline

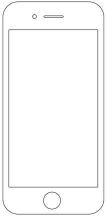
- Background: Core Model and Implementation
- Work 1: Rethinking Model Training
- **Work 2: Rethinking Generation (Learning Neural Templates)**
- Challenges: Text Generation and Deep Learning

Talk about Data

x

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

f_θ

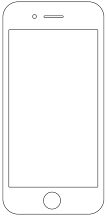


Talk about Data

x

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

f_θ



$y_{1:T}$

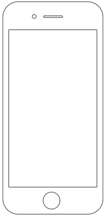
Fitzbillies is a coffee shop providing Chinese food in the moderate price range . It is located in the city centre . Its customer rating is 3 out of 5 .

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

θ

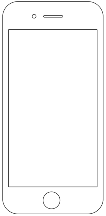


Talking About Data

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

θ



$y_{1:T}$

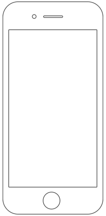
Frederick Parker-Rhodes (21 November 1914 2 March 1987) was an English linguist, plant pathologist, computer scientist, mathematician, mystic, and mycologist.

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

θ



$y^*_{1:T}$

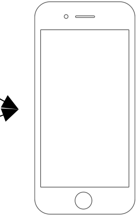
Frederick Parker-Rhodes (21 November 1914 2 March 1987) was an English mycology and plant pathology, mathematics at the University of UK.

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

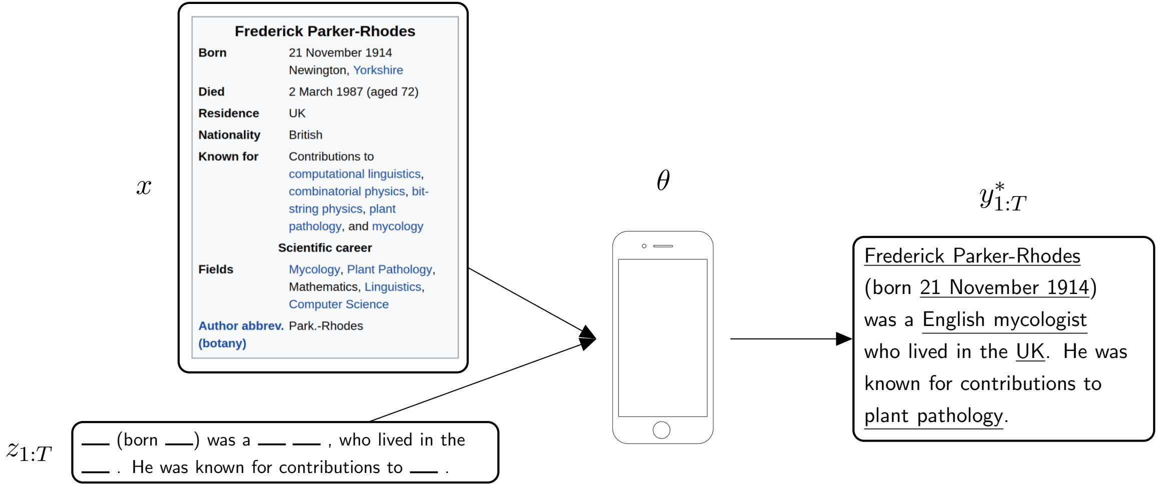
θ



$z_{1:T}$

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

Talking About Data



Arguments for Templated Generation

Guarantees about the quality, in particular,

1. Interpretable in its factual content.
2. Controllable in terms of style.

Goal: Can we achieve this with a deep-learning based system?

Technical Approach: Deep Latent-Variable Models

Expose specific choices as latent variables z.

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

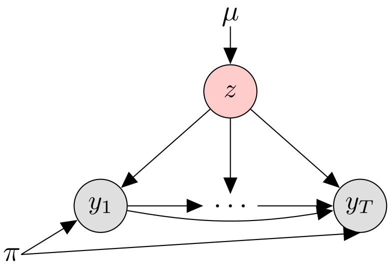
- x, y, θ as before, *what to talk about / how to say it*
- z is a collection of latent variables

Preliminary Model 1: Sentence Clusters

Generative process:

1. Draw cluster $z \in \{1, \dots, Z\}$ from a Categorical.
2. Draw words $y_{1:T}$ from decoder RNN with parameters π_z.

$$p(y, z \mid x; \theta) = \mu_z \times \mathrm{RNN}(y_{1:T}; \pi_z)$$



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

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

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

Preliminary Model 2: Neural Copy Models

Generative process:

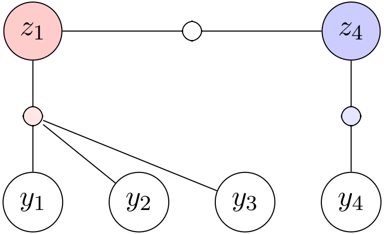
1. Draw copy switch $z \in \{0, 1\}$ from a Bernoulii.
2. Draw words $y_{1:T}$ from decoder RNN where
   - If $z = 0$, let the model generate a new word.
   - If $z = 1$, let the model copy a word from the source.

Example:

<u>Frederick Parker-Rhodes</u> (born <u>21 November 1914</u>) was a English <u>linguist</u> ...

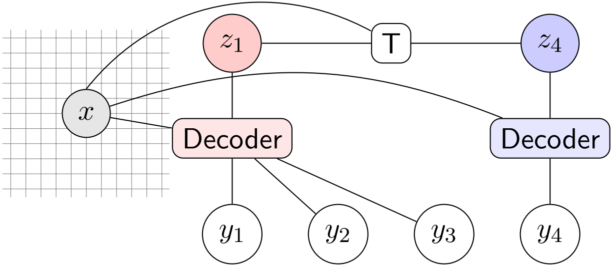
Classical Model: Hidden Semi-Markov Model

- Hidden Markov Model: discrete latent states with single emissions (e.g. words).
- Extension: discrete latent states produce multiple emissions (e.g. phrases).
- Parameterized with *transition*, *emission*, and *length* distributions.



A Deep Hidden Semi-Markov Model

- Employ HSMM as a conditional latent variable language model, $p(y_1, \dots, y_T, z \mid x)$.
- Transition Distribution: neural network between clusters.
- Emission Distribution: Encoder-Decoder+Copy, specialized per cluster $\{1, \dots, Z\}$.



Technical Methodology: Training Model

Fit model by minimizing negative log-marginal likelihood on training data.

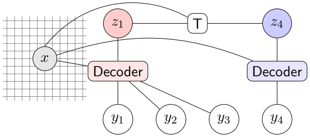
$$\mathcal{L}(\theta) = -\log \sum_{z_{1:T}} p(\hat{y}_{1:T}, z_{1:T} \mid x; \theta)$$

- Dynamic programming to efficiently compute HSMM forward algorithm for sum
- Backpropagation with autograd, sum computation is exact.

However, this just gives another score model $f_\theta(y_{1:T}, x)$. Want templates.

From Neural HSMM to Templates

Extract "templates" by finding most common, best sequences of training sentences.



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

The Wrestlers is a coffee shop that serves English . . .

$$\Downarrow z^*$$

The Wrestlers | is a | coffee shop | that serves | English . . .

Example Templates: Wikipedia

Sentences grouped by the same $z^*_{1:T}$ and their splits.

1.

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

2.

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

3.

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

4.

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

5.

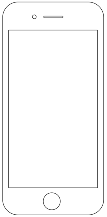
aftab ahmed	is a	member of the	knesset	.
anderson da silva	is a former	party member of the	scottish parliament	
david jones	is a female	recipient of the	fc lokomotiv liski	
…	…	…	…	

Neural Template Generation Approach

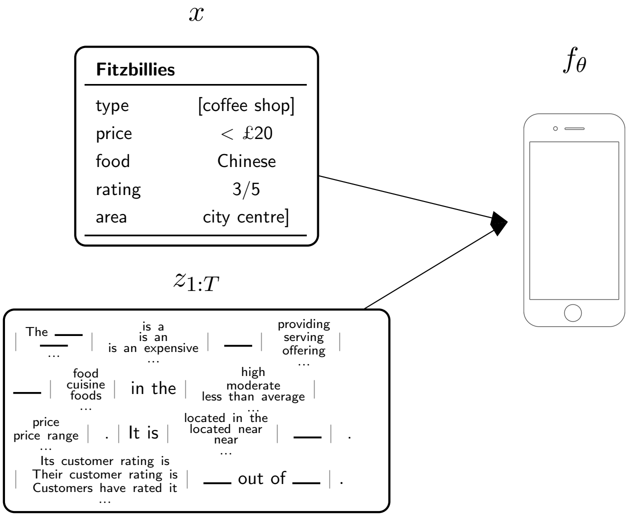
x

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

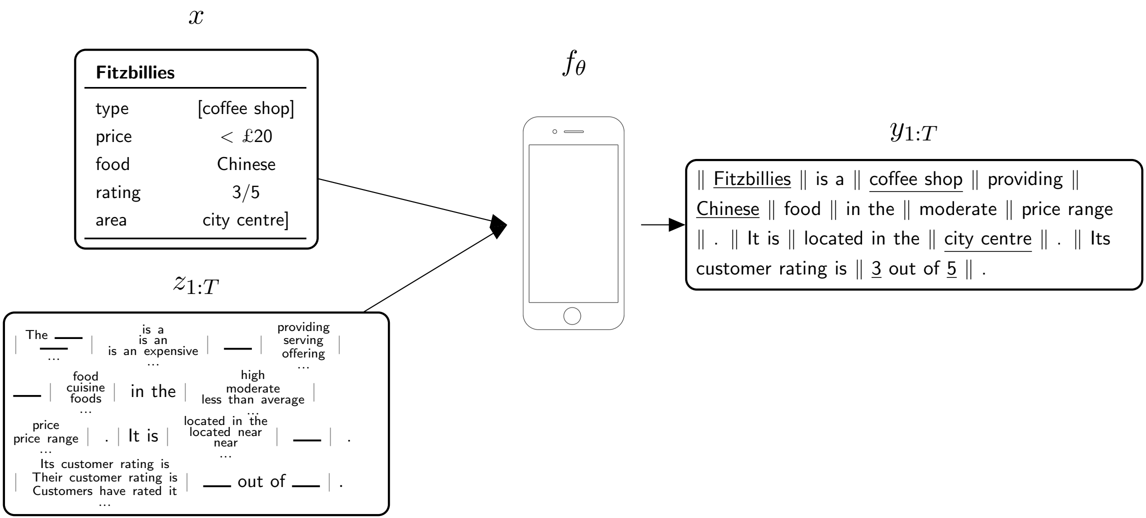
f_θ



Neural Template Generation Approach



Neural Template Generation Approach



Issue 1: Interpretability

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.

Issue 2: Controllability

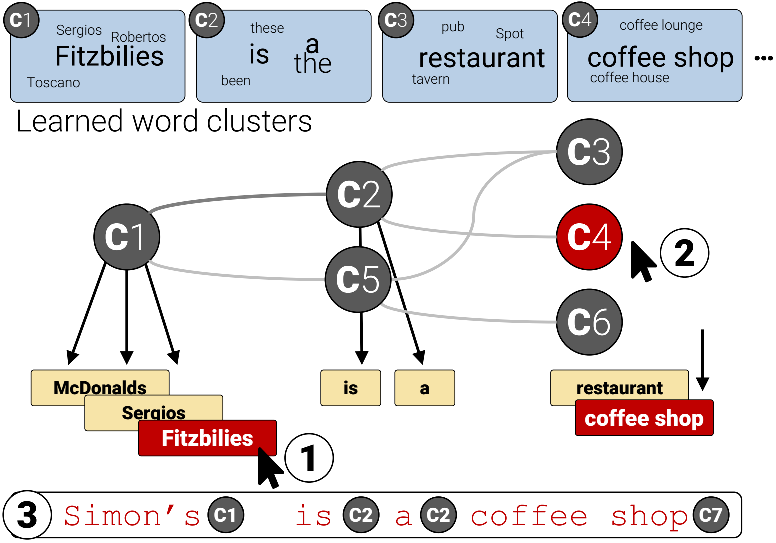
The Golden Palace
name[The Golden Palace], type[coffee shop], food[Chinese], priceRange[cheap] custRating[5 out of 5], area[city centre],
1. The Golden Palace is a coffee shop located in the city centre. 2. In the city centre is a cheap Chinese coffee shop called The Golden Palace. 3. The Golden Palace 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.

Results

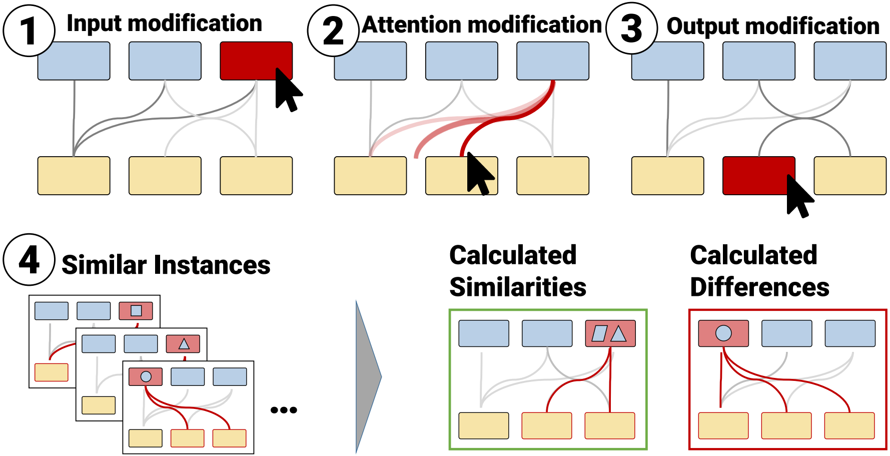
	BLEU	NIST
Test		
Substitution	43.78	6.88
Neural Template	56.72	7.63
Full Neural Model	65.93	8.59

	BLEU	NIST	ROUGE-4
Conditional KN-LM	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	33.8	7.51	28.2

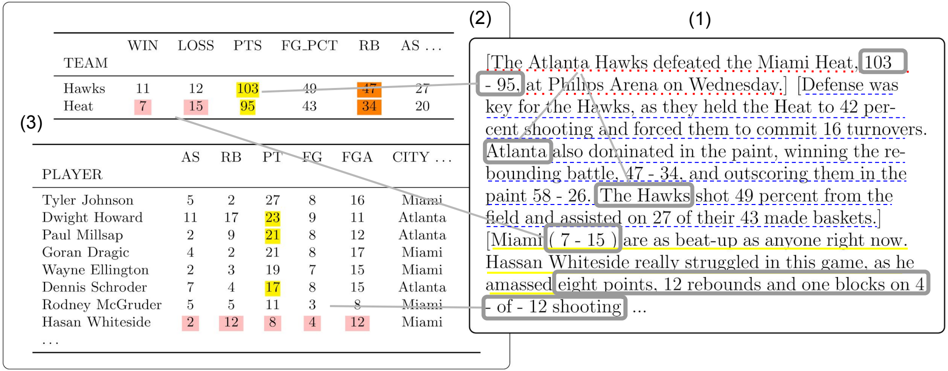
Controllable Interactive Deep Learning Systems



Another Application: Understanding Model Selection



Long-Form Generation with Explicit Reasoning



TEAM	WIN	LOSS	PTS	FG_PCT	RB	AS ...
Hawks	11	12	103	49	47	27
Heat	7	15	95	43	34	20

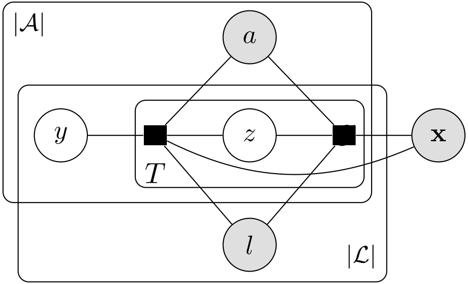
PLAYER	AS	RB	PT	FG	FGA	CITY ...
Tyler Johnson	5	2	27	8	16	Miami
Dwight Howard	11	17	23	9	11	Atlanta
Paul Millsap	2	9	21	8	12	Atlanta
Goran Dragic	4	2	21	8	17	Miami
Wayne Ellington	2	3	19	7	15	Miami
Dennis Schroder	7	4	17	8	15	Atlanta
Rodney McGruder	5	5	11	3	8	Miami
Hasan Whiteside	2	12	8	4	12	Miami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

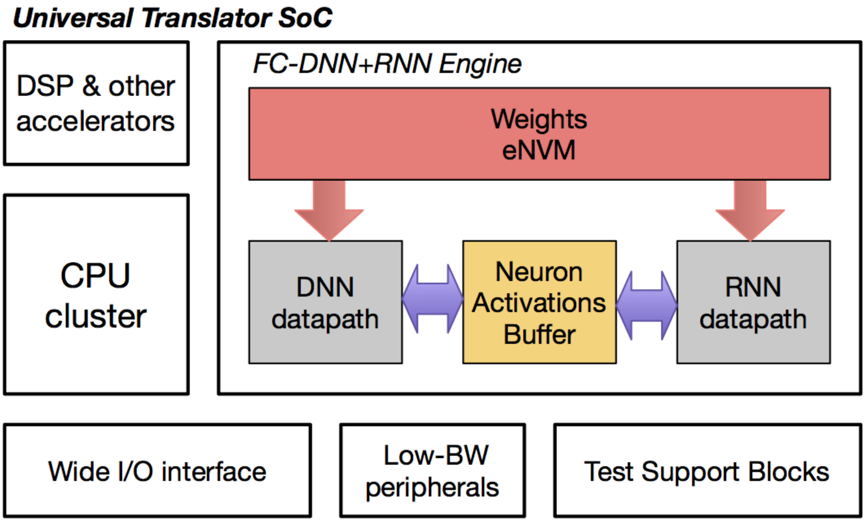
1. Discourse-aware structure in generation
2. Explicit Linking and coreference
3. Aggregation of factual information before generation

Talk Outline

1. Background: Core Model and Implementation
2. Work 1: Rethinking Model Training (*Beam Search Optimization*)
3. Work 2: Rethinking Generation (*Learning Neural Templates*)
4. **Future Challenges Beyond Text Generation**

Simpler and Cleaner Open-Research


$|\mathcal{A}|$
a
y
z
$\mathbf{x}$
T
l
$|\mathcal{L}|$


Universal Translator SoC
DSP & other accelerators
CPU cluster
FC-DNN+RNN Engine
Weights eNVM
DNN datapath
Neuron Activations Buffer
RNN datapath
Wide I/O interface
Low-BW peripherals
Test Support Blocks

Challenges in Discrete Deep Learning

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*.