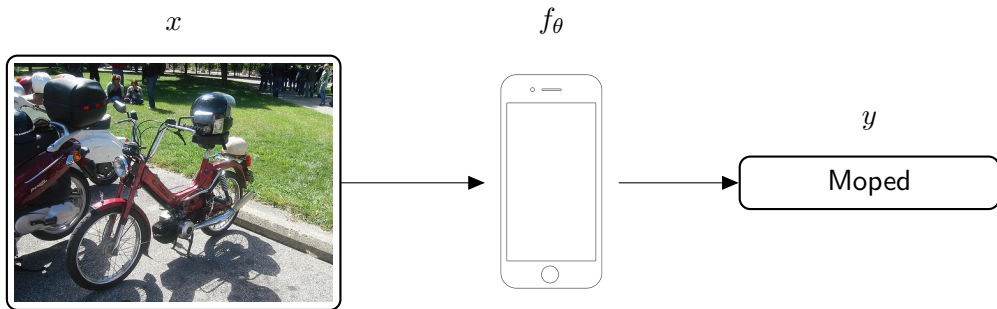


Learning How to Say It: Language Generation and Deep Learning

Alexander M Rush

Machine Learning for Multiclass Classification

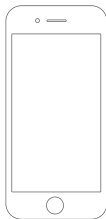


Machine Learning for Text Generation: Translation

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

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

- Input \mathbf{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

Training and Evaluation

Training θ :

- Data consists of paired examples (x, \hat{y}) , *how people say it*
- Typically as large as 100,000 to 10 million examples.

Training and Evaluation

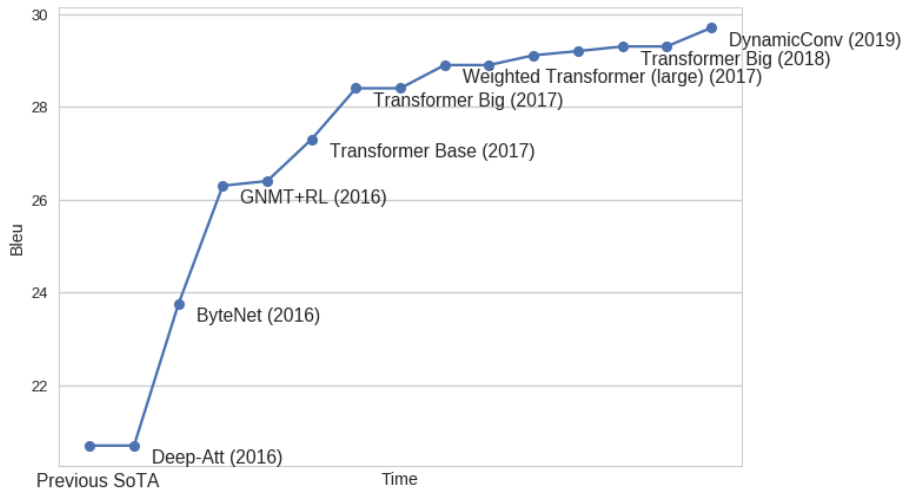
Training θ :

- Data consists of paired examples (x, \hat{y}) , *how people say it*
- Typically as large as 100,000 to 10 million examples.

Evaluation:

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

Machine Translation Performance

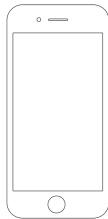


Sentence Summarization

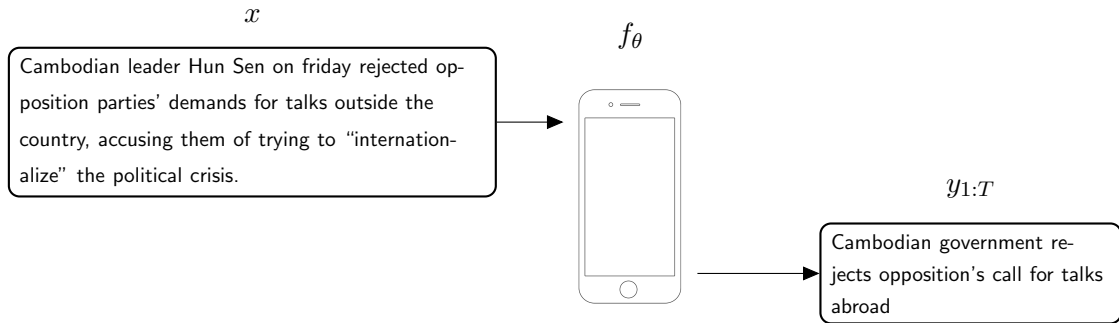
x

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



Sep 13, 3:17 PM EDT

GERMANY IMPLEMENTS TEMPORARY BORDER CHECKS TO LIMIT MIGRANTS

BY GEIR MOULSON AND SHAWN POGATCHNIK
ASSOCIATED PRESS

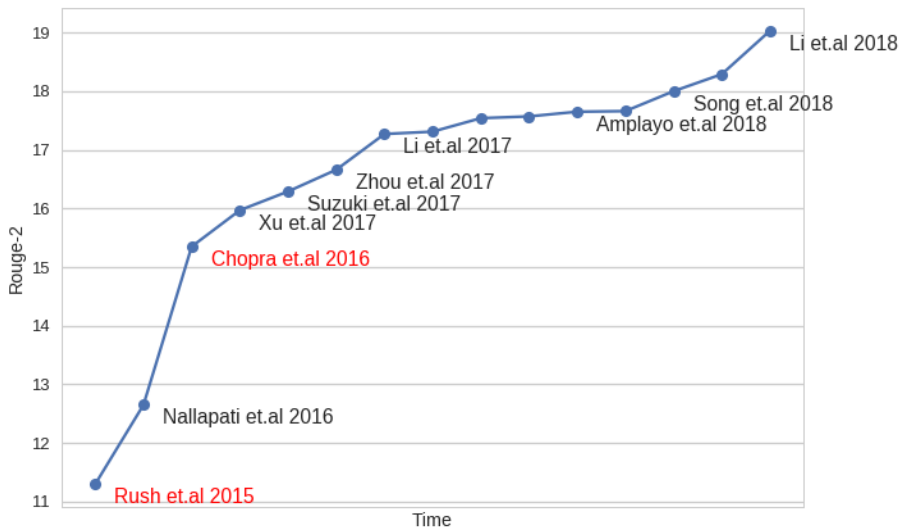
BERLIN (AP) -- Germany introduced temporary border controls Sunday to stem the tide of thousands of refugees streaming across its frontier, sending a clear message to its European partners that it needs more help with an influx that is straining its ability to cope.

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



AP Photo/Kay Nietfeld

Sentence Summarization Performance



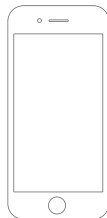
Document Summary

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
...						

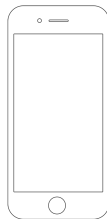


The Atlanta Hawks defeated the Miami Heat, 103 - 95, at Philips Arena on Wednesday. Atlanta was in desperate need of a win and they were able to take care of a short-handed Miami team here. Defense was key for the Hawks, as they held the Heat to 42 percent shooting and forced them to commit 16 turnovers.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

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



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


Convert images to LaTeX

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



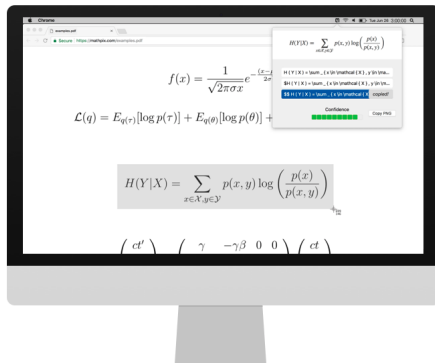
MacOS



Windows



Ubuntu



Outline

Goal

Learn How to Say It

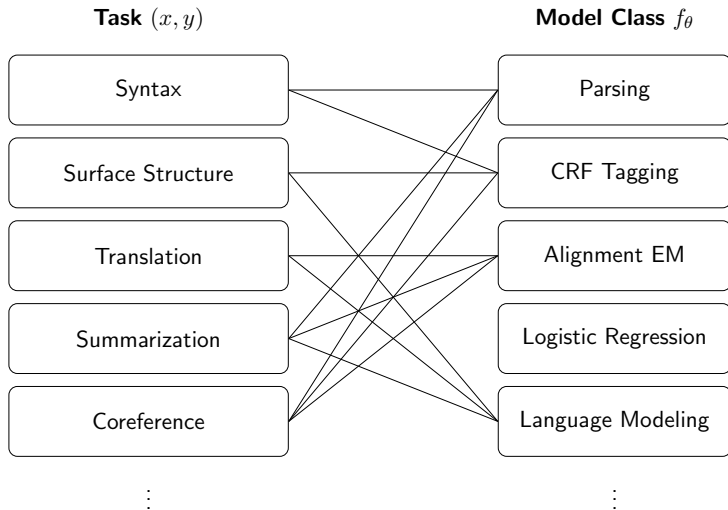
Outline

Goal

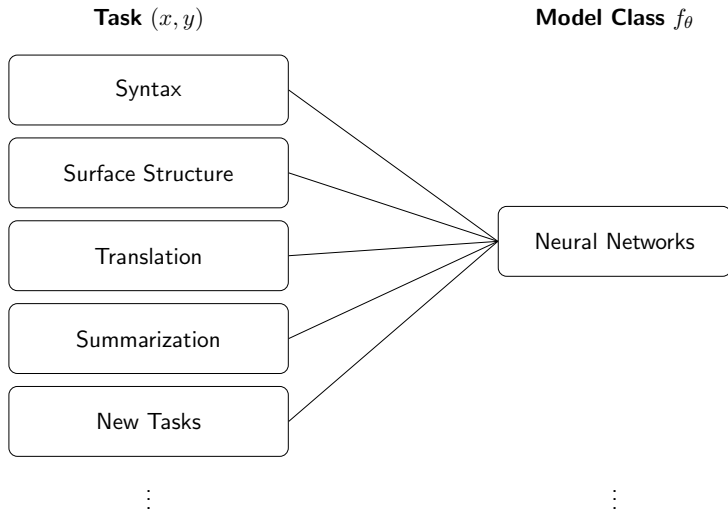
Learn How to Say It

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

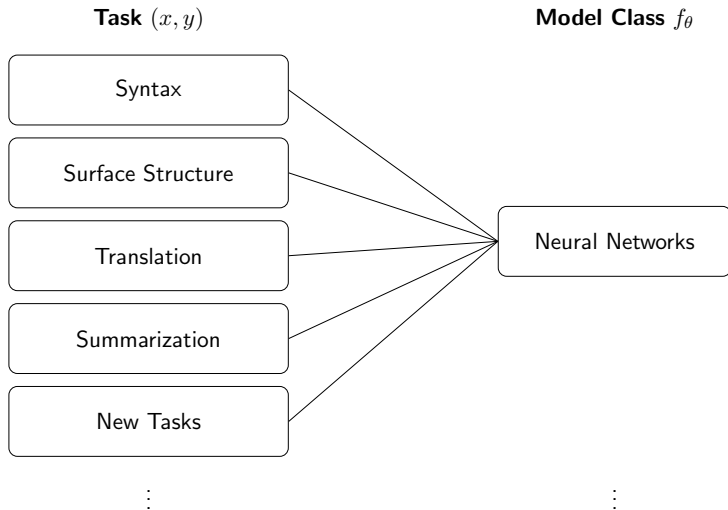
State-of-the-Art Natural Language Processing, circa 2009



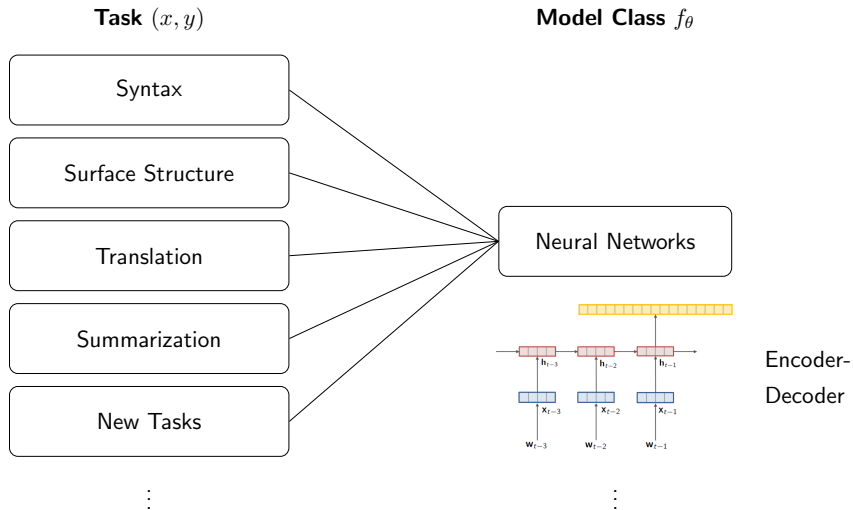
State-of-the-Art Natural Language Processing, circa 2019



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



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



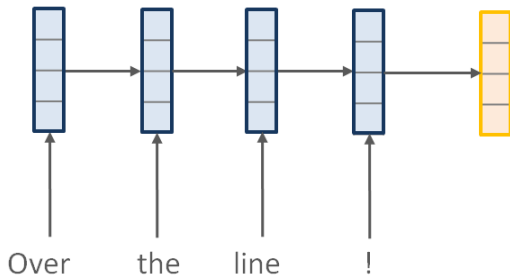
Encoder-Decoder

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

Over the line !

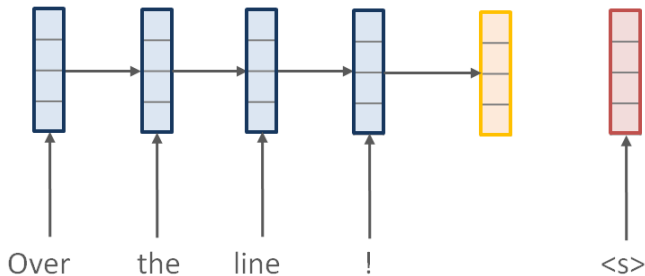
Encoder-Decoder

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



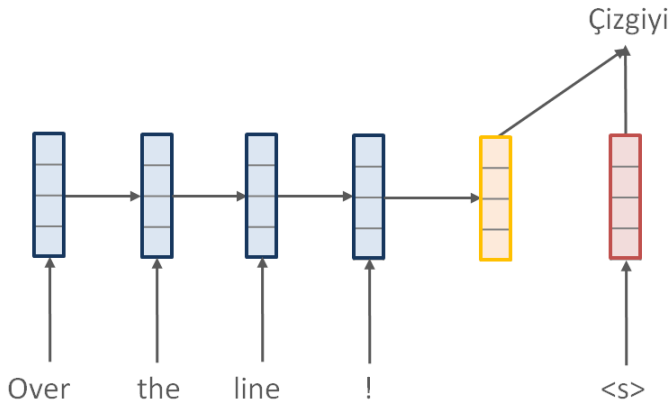
Encoder-Decoder

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



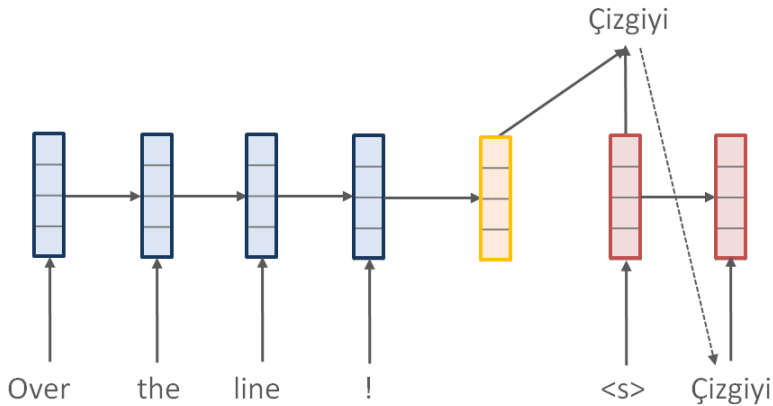
Encoder-Decoder

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



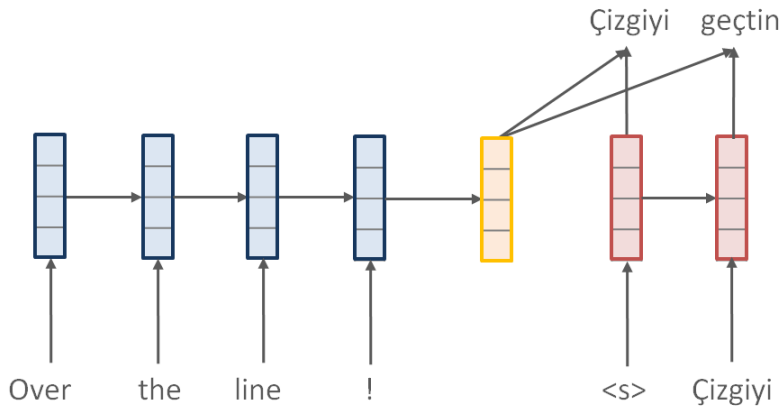
Encoder-Decoder

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



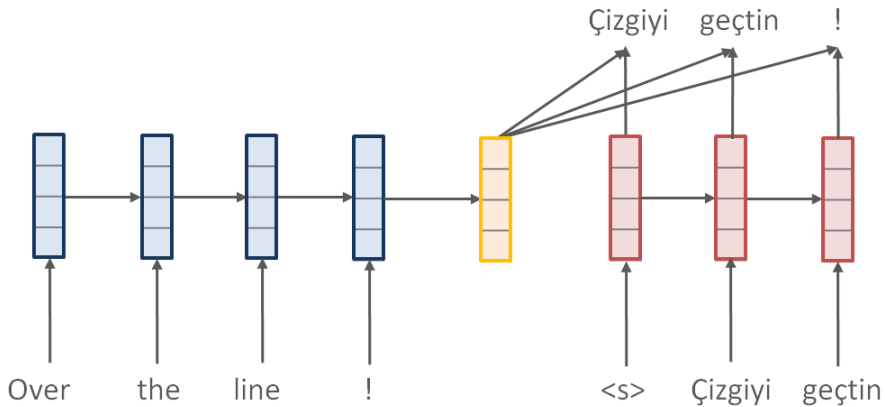
Encoder-Decoder

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



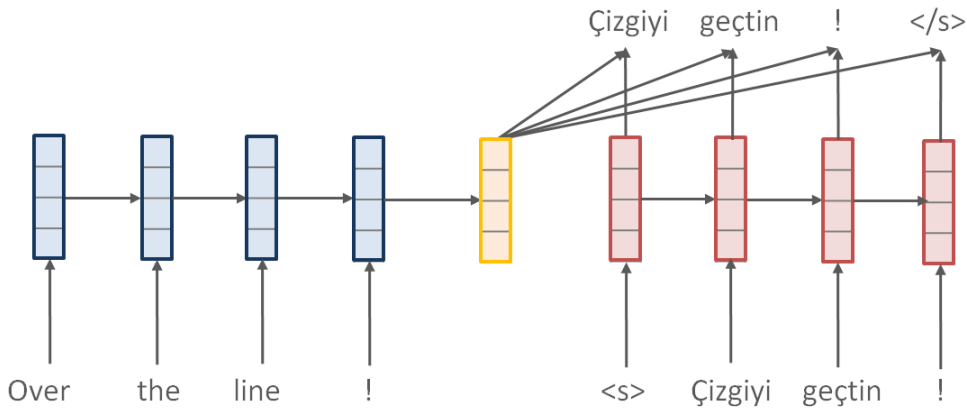
Encoder-Decoder

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



Encoder-Decoder

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



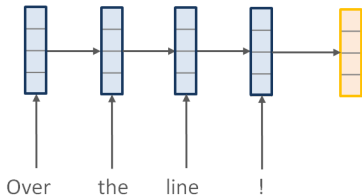
Encoder-Decoder

Encoder:

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

Context:

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



Encoder-Decoder

Encoder:

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

Context:

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

Decoder:

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

Scoring function:

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

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

Scoring Model Example

Decoder:

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

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

- Vocabulary: () 0 1 2 3 4

$$f(\text{"0 ((2) (((4 4 4) 3) ..."}, x)$$

>

$$f(\text{"0) (3)) (() 2 ()) 2" ...}, x)$$

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



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](#)

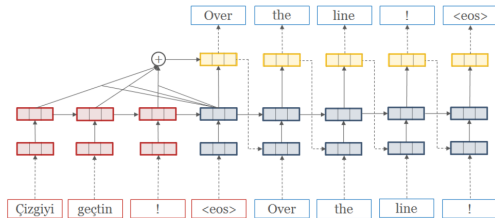
[FAQ](#)

[About](#)

[Documentation](#)

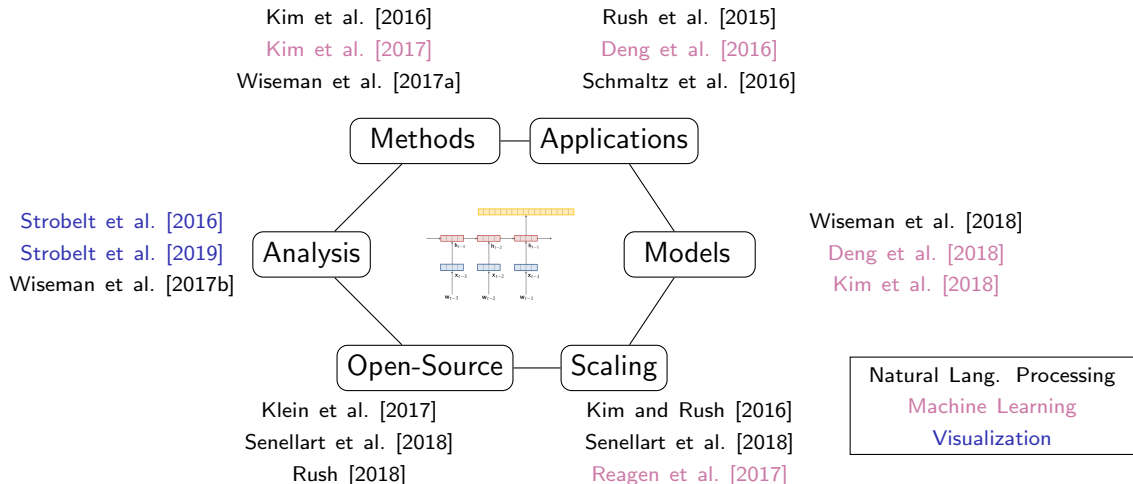
Home

OpenNMT is an 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



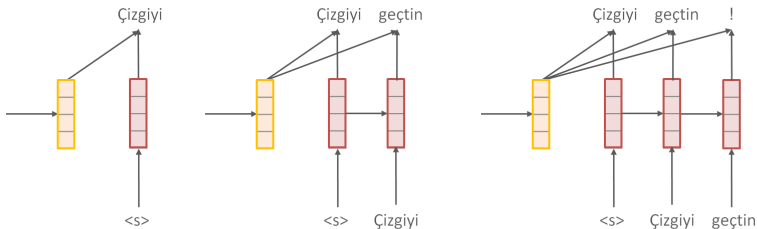
Outline

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

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

Baseline: Training Encoder-Decoder

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

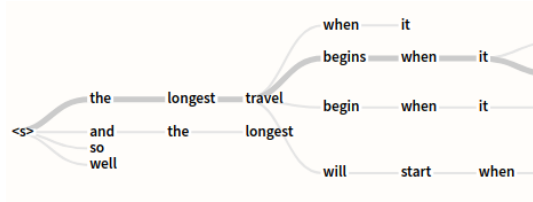
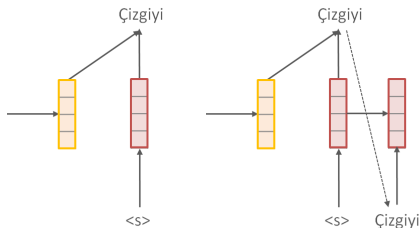


Training loss is identical to **multiclass classification**,

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

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

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

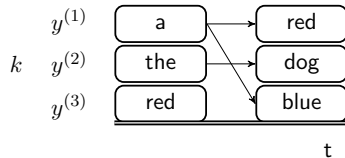
k

$y^{(1)}$	a
$y^{(2)}$	the
$y^{(3)}$	red

t

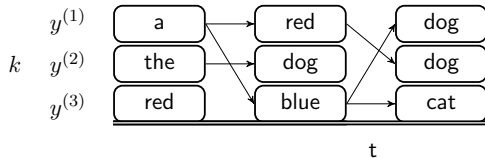
Standard Heuristic Method: Beam Search

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



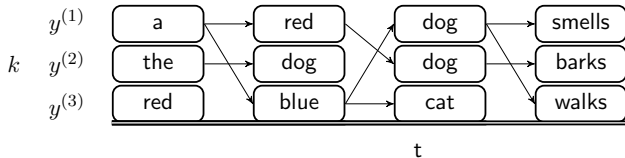
Standard Heuristic Method: Beam Search

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



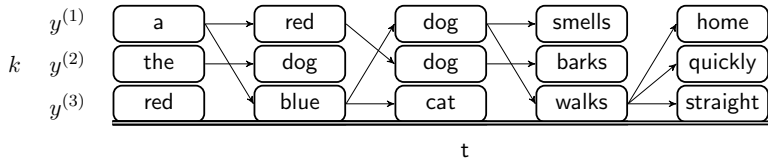
Standard Heuristic Method: Beam Search

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



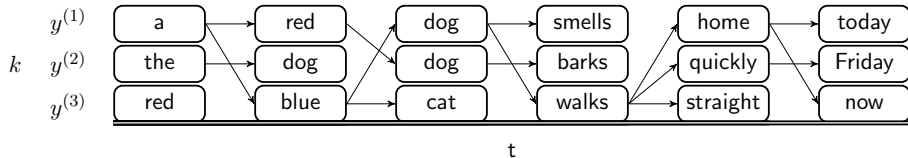
Standard Heuristic Method: Beam Search

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



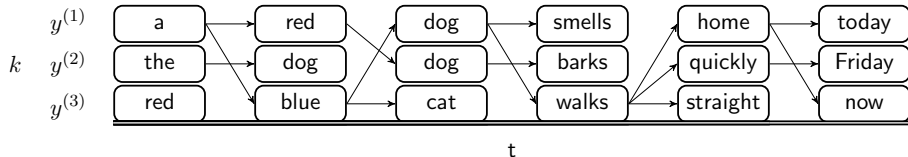
Standard Heuristic Method: Beam Search

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



Standard Heuristic Method: Beam Search

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



- 1 For each k in beam, example all possible next words y_t ,

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

- 2 Prune $(K \times \text{vocab})$ expansions to top K ,

$$y_{1:t}^{(1:K)} \leftarrow K \arg \max_{y_t, k} f_{\theta}(\langle y_t, y_{1:t}^{(k)} \rangle, x)$$

Generation Training Issues

Multiclass Training \Rightarrow Structured Generation ?

Generation Training Issues

Multiclass Training \Rightarrow Structured Generation ?

① Exposure Bias

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

Generation Training Issues

Multiclass Training \Rightarrow Structured Generation ?

① Exposure Bias

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

② Label Bias

- Score is locally multiclass, but want to compare entire sequences.

Generation Training Issues

Multiclass Training \Rightarrow Structured Generation ?

① Exposure Bias

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

② Label Bias

- Score is locally multiclass, but want to compare entire sequences.

③ Metric Bias

- Training error is class accuracy, but evaluation uses n-gram match.

Generation Training Issues

Multiclass Training \Rightarrow Structured Generation ?

① Exposure Bias

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

② Label Bias

- Score is locally multiclass, but want to compare entire sequences.

③ Metric Bias

- Training error is class accuracy, but evaluation uses n-gram match.

Strategy: Modify model and training to target these issues.

Modification 1: Beam Search at Training

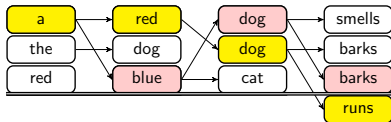
Fix: Exposure Bias

- 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. compare true sequence $\hat{y}_{1:t}$ to $y_{1:t}^{(K)}$ worst in beam

Modification 2: Global Scoring Function

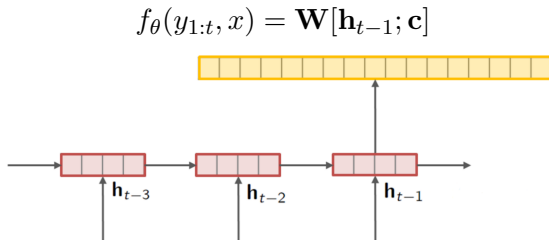
Fix: Label Bias

- Use a global sequence scoring function.

Modification 2: Global Scoring Function

Fix: Label Bias

- Use a global sequence scoring function.



- 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.

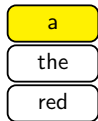
Extension: Training with Hard Constraints

Beam Search Optimization supports hard constraints at training.

Example: Code generation with a known grammar,

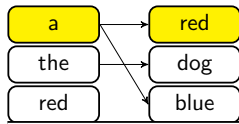
```
{ \cal K } ^ { L } ( \sigma = 2 ) = \left( \begin{array}{l}
{ c c } { - \frac { d ^ { 2 } } { d x ^ { 2 } } + 4 - \frac { 3 } { 3 } { \operatorname { c o s h } ^ { 2 } x } } \& { \frac { 3 } { 3 } { d x ^ { 2 } } } { \frac { 3 } { 3 } { \operatorname { c o s h } ^ { 2 } x } } \& { - \frac { d ^ {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

Beam Search Optimization: Training Example



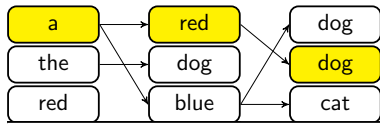
- **True:** ground-truth sequence $\hat{y}_{1:t}$

Beam Search Optimization: Training Example



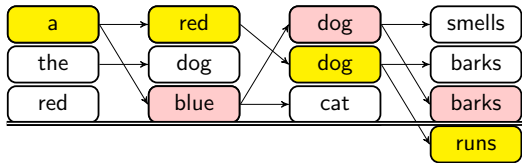
- **True:** ground-truth sequence $\hat{y}_{1:t}$

Beam Search Optimization: Training Example



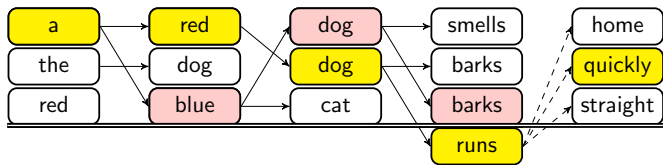
- **True:** ground-truth sequence $\hat{y}_{1:t}$

Beam Search Optimization: Training Example



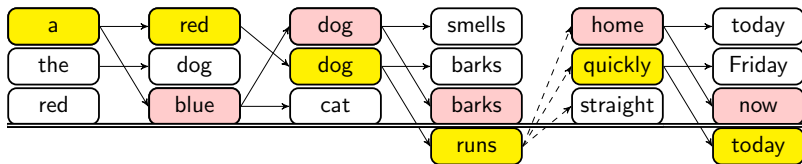
- **True:** ground-truth sequence $\hat{y}_{1:t}$
- **Predicted:** lowest-scoring violating sequence $y_{1:t}^{(K)}$

Beam Search Optimization: Training Example



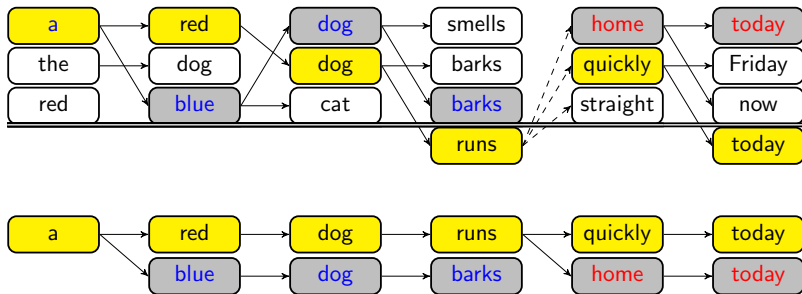
- **True:** ground-truth sequence $\hat{y}_{1:t}$
- **Predicted:** lowest-scoring violating sequence $y_{1:t}^{(K)}$

Beam Search Optimization: Training Example



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

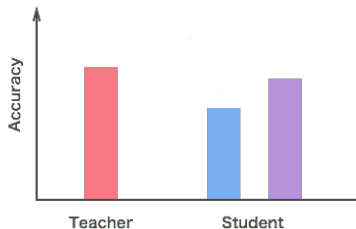
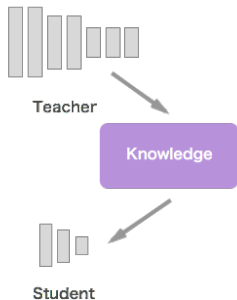
Parameter Updates: Structured Backpropagation



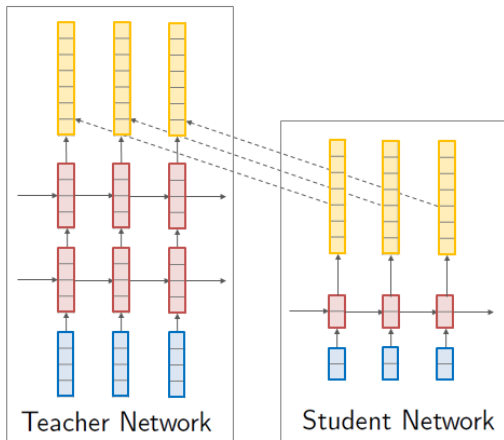
Results

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

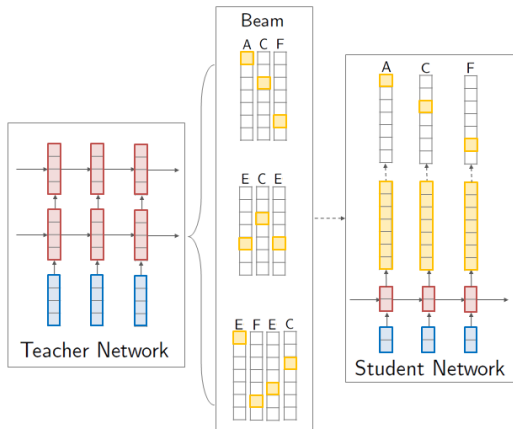
Knowledge Distillation: Train a *student* model to learn from a *teacher* model.



Baseline: Word-Level Knowledge Distillation (Multiclass)



Sequence-Level Knowledge Distillation



Results: WMT English \rightarrow German Translation

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

Results: WMT English \rightarrow German Translation

Model	$\text{BLEU}_{K=1}$	$\Delta_{K=1}$	$\text{BLEU}_{K=5}$	$\Delta_{K=5}$
<hr/>				
4 \times 1000				
Teacher	17.7	—	19.5	—
<hr/>				
2 \times 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 _{$K=5$}	$\Delta_{K=5}$
<hr/>				
4 \times 1000				
Teacher	17.7	—	19.5	—
<hr/>				
2 \times 500				
Student	14.7	—	17.6	—
Word-KD	15.4	+0.7	17.7	+0.1
Seq-KD	18.9	+4.2	19.3	+1.7
<hr/>				

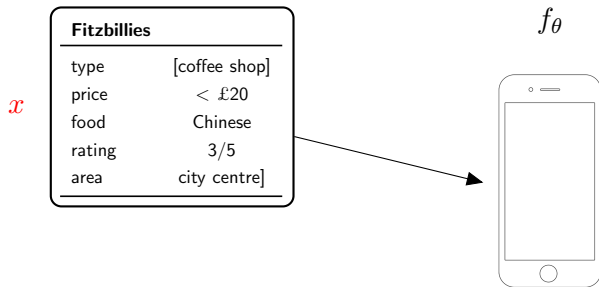
Scaling Translation Models

Outline

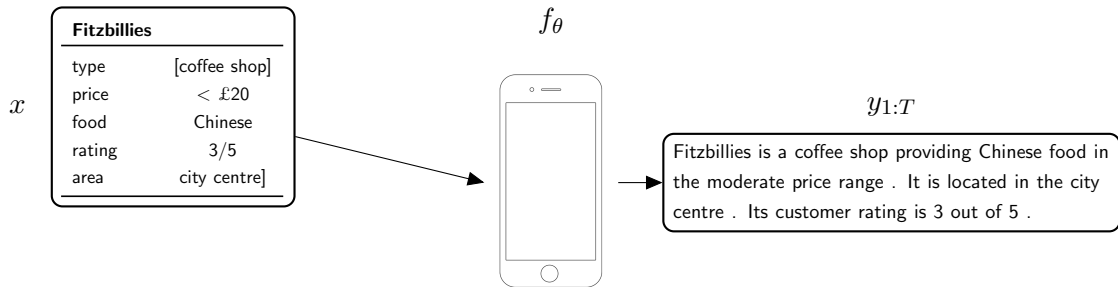
- Background: Core Model and Implementation
- Work 1: Rethinking Model Training
- **Work 2: Rethinking Generation (Learning Neural Templates)**
- Challenges: Text Generation and Deep Learning

Can we learn interpretable and controllable target text generation models?

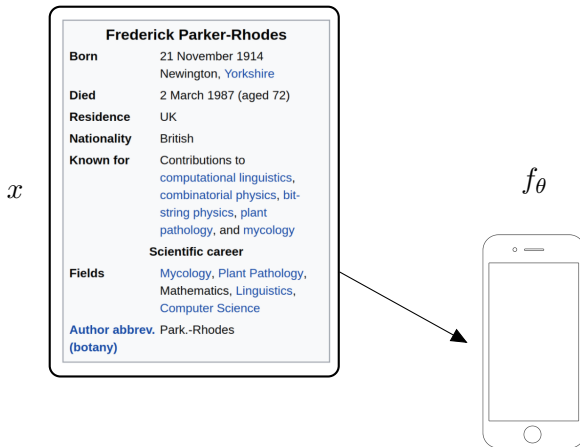
Talk about Data



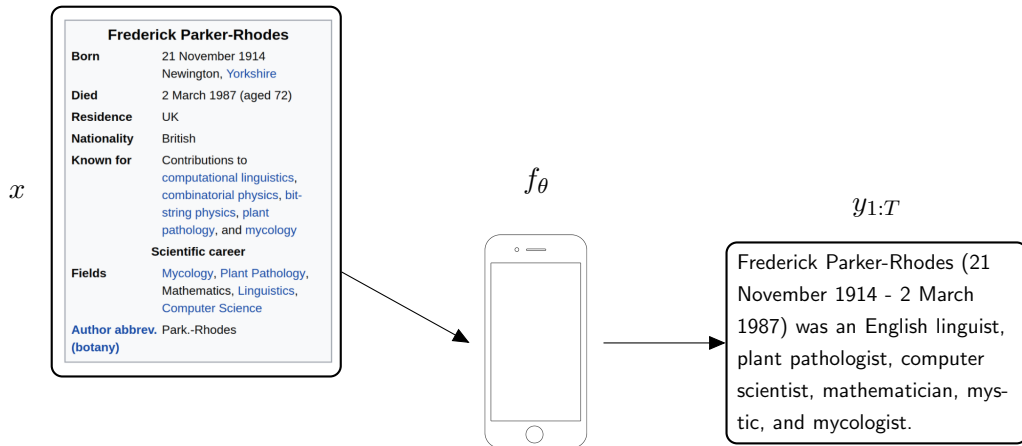
Talk about Data



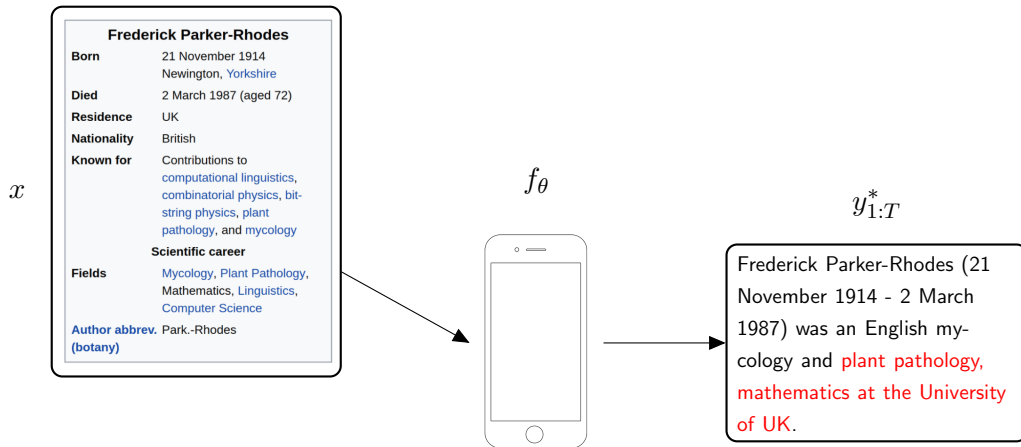
Talking About Data



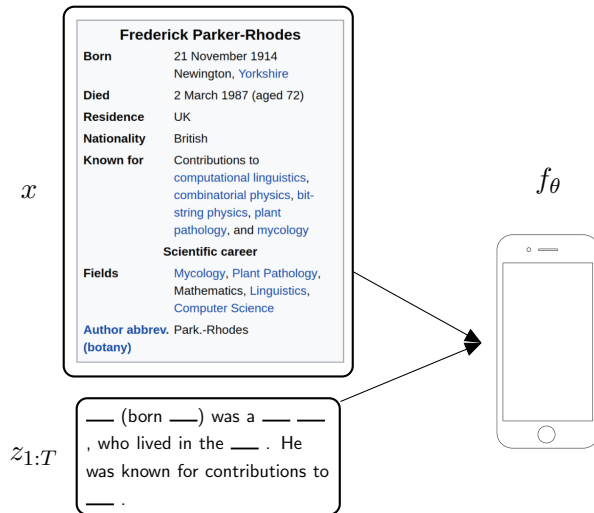
Talking About Data



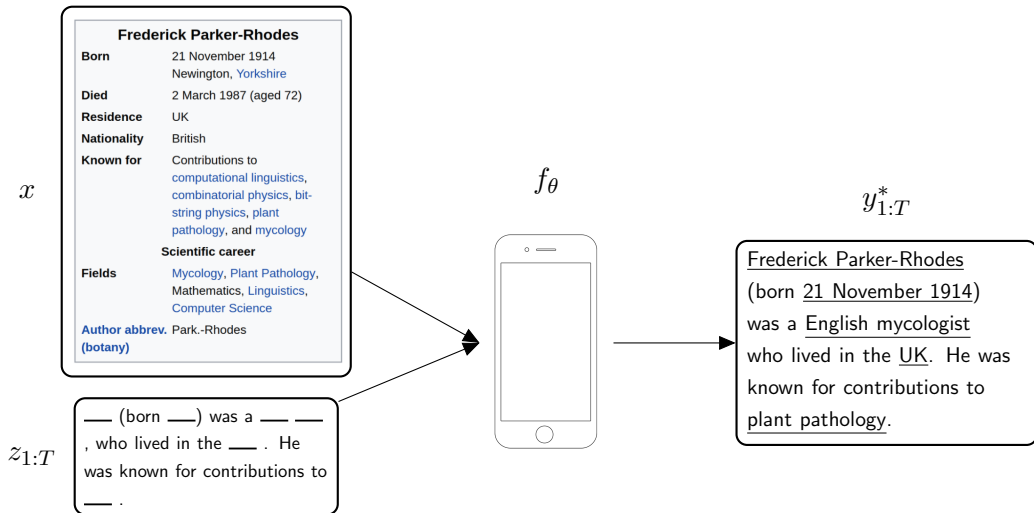
Talking About Data



Talking About Data



Talking About Data



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 with a deep learning system?

Technical Approach: Deep Latent-Variable Models

Expose specific choices as latent variables z .

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

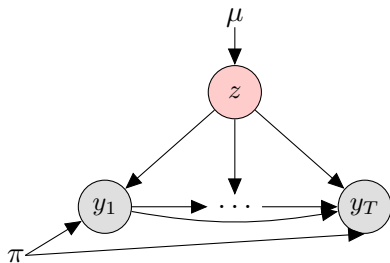
- x, y, θ as before, *what to talk about, how to say it*
- z is a collection of problem-specific latent variables

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 .

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



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: Deep Copy Models

Generative process:

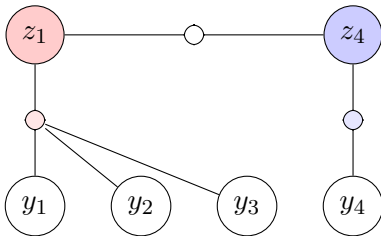
- ① Draw copy switch $z_t \in \{0, 1\}$ from a Bernoulli.
- ② Draw words y_t where
 - If $z = 0$, generate a new word from the decoder.
 - If $z = 1$, copy a word from the source data.

Example:

Frederick Parker-Rhodes (born 21 November 1914) was a English linguist ...

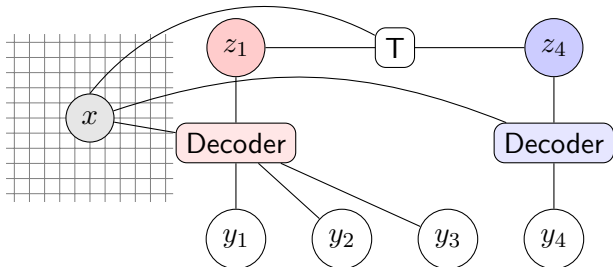
Classical Model: Hidden Semi-Markov Model

- Extension of Hidden Markov Model where discrete latent states produce multiple emissions (e.g. phrases).
- Parameterized with *transition* and *emission* 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.

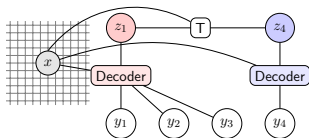
$$\mathcal{L}(\theta) = -\log \sum_{z_{1:T}} p(\hat{y}_{1:T}, z_{1:T} \mid x; \theta)$$

- Requires summing over all clusters and segmentation.
- Approach: backprop through HSMM dynamic programming algorithm, exact gradient.
- Gives a new definition for $f_{\theta}(y_{1:T}, x)$

However, this model alone does not give templates.

From Neural HSMM to Templates

Extract templates by finding most common, best latent sequences from training.



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

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 kneset
 | 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 american 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 kneset
 | 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

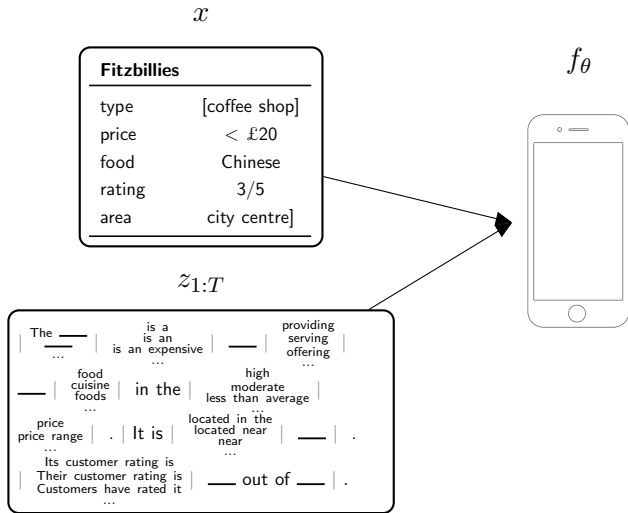
x

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

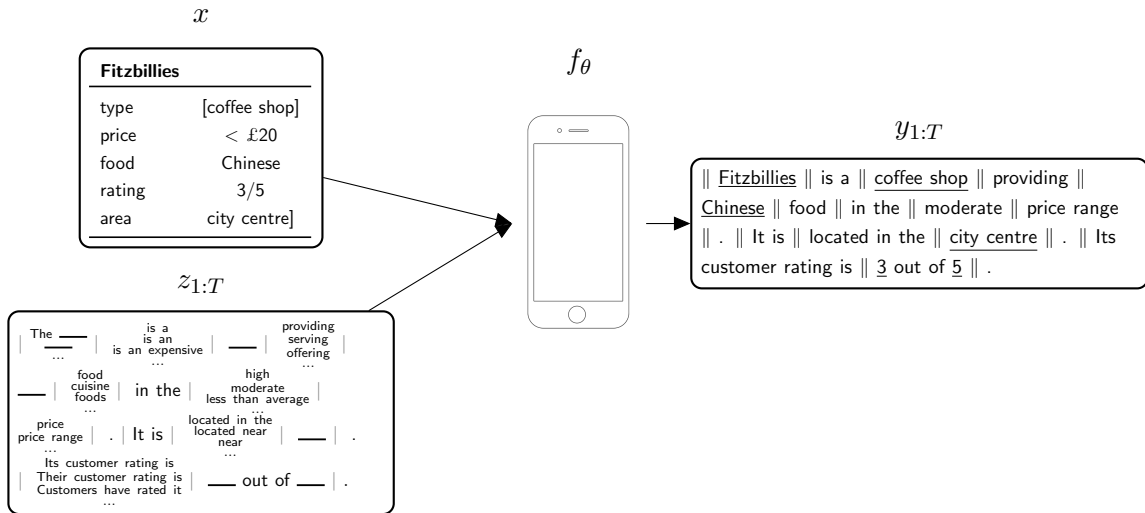
f_{θ}



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],

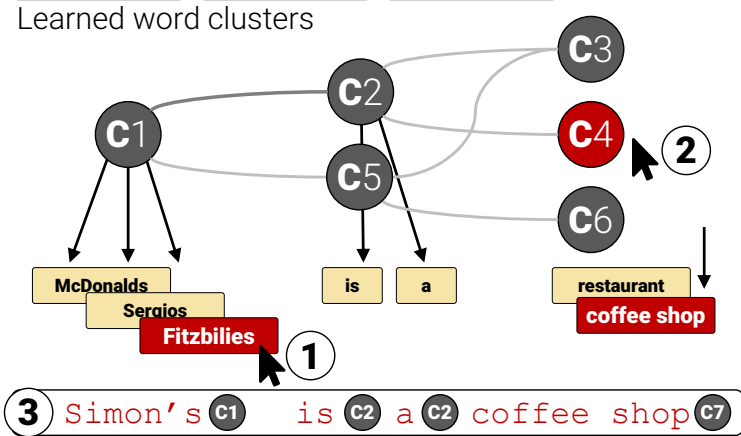
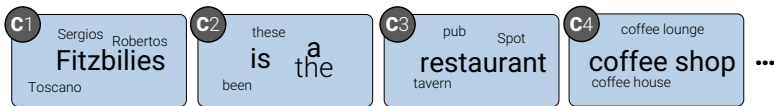
1. The Golden Palace is a coffee shop located in the city centre.
 2. In the city centre is a cheap Chinese coffee shop called The Golden Palace.
 3. The Golden Palace is a Chinese coffee shop.
 4. The Golden Palace is a Chinese coffee shop with a customer rating of 5 out of 5.
 5. The Golden Palace 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

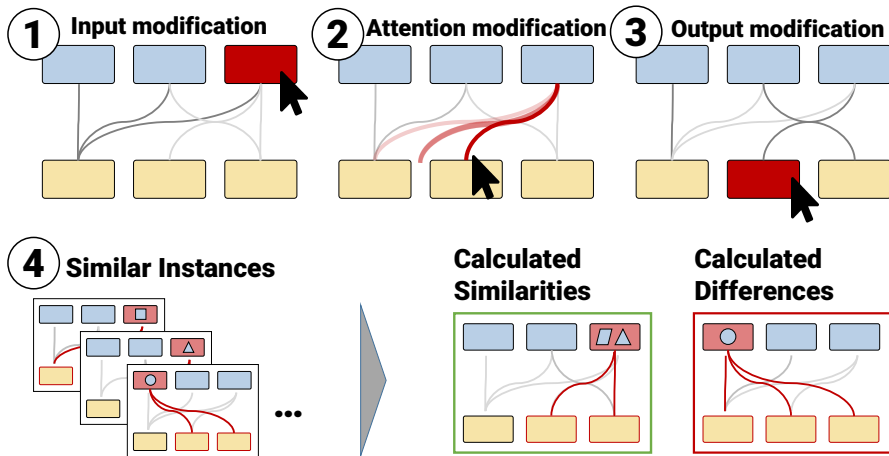
	BLEU	NIST
Test		
Substitution	43.78	6.88
Neural Template	56.72	7.63
Full Neural Model	65.93	8.59

	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



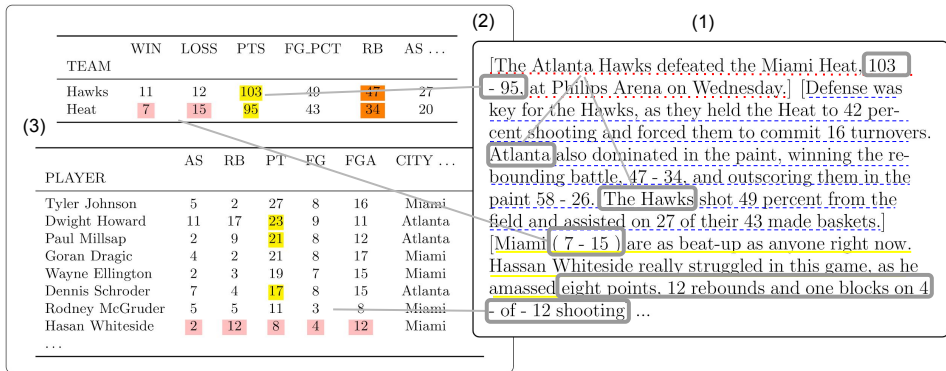
Another Application: Understanding Model Selection



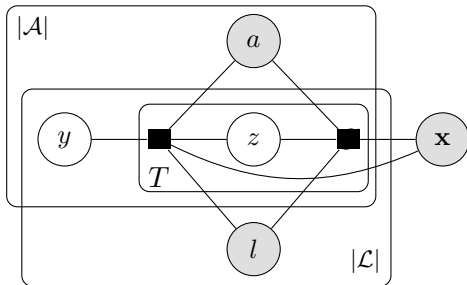
Talk Outline

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

Long-Form Generation with Explicit Reasoning

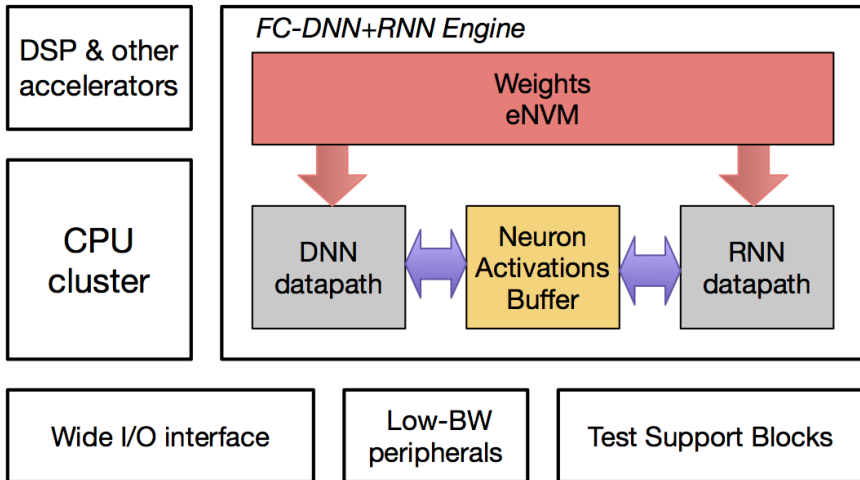


- 1 Discourse-aware structure in generation
- 2 Explicit Linking and coreference
- 3 Aggregation of factual information before generation



Hardware Co-Design for Generation and Understanding print

Universal Translator SoC



Challenges in Discrete Deep Learning

Thanks

<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>

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](https://arxiv.org/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*.