Learning How to Say It:
Language Generation and Deep Learning

Alexander M Rush

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 Directions

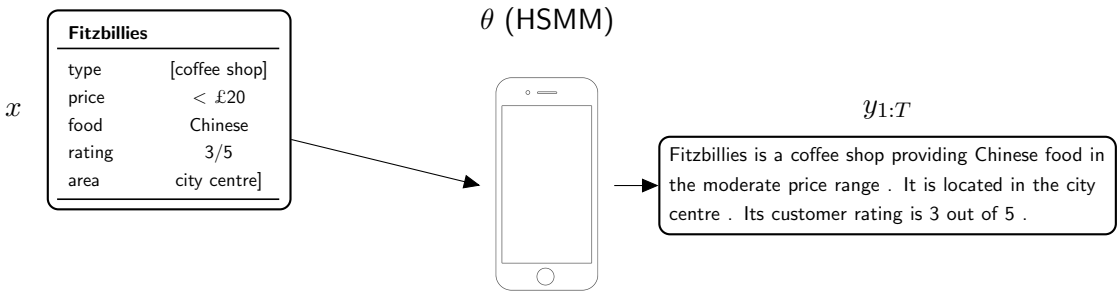
Talk about Data

x

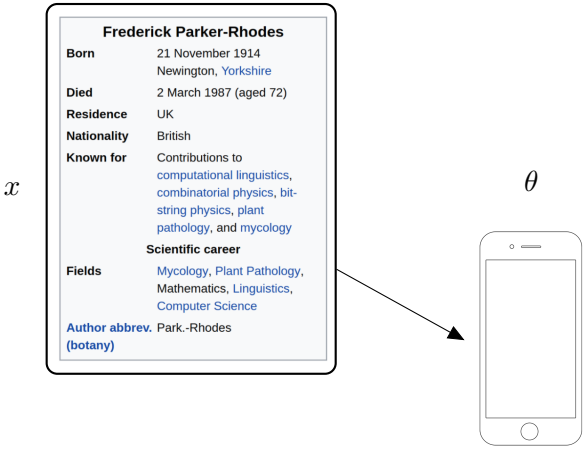
Fitzbillies	
type	[coffee shop]
price	$<$ £20
food	Chinese
rating	3/5
area	city centre]

θ (HSMM)

Talk about Data



Talking About Data



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

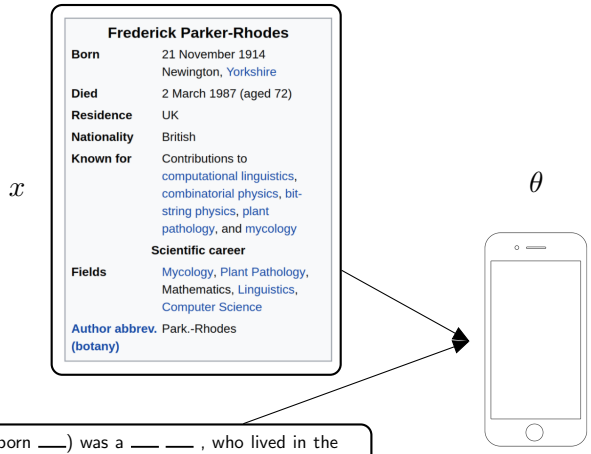
Born	21 November 1914 Newington, Yorkshire
Died	2 March 1987 (aged 72)
Residence	UK
Nationality	British
Known for	Contributions to computational linguistics, combinatorial physics, bit-string physics, plant pathology, and mycology
	Scientific career
Fields	Mycology, Plant Pathology, Mathematics, Linguistics, 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



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

$z_{1:T}$

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

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.

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

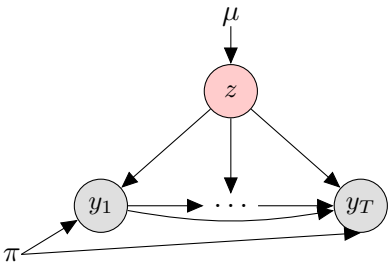
- x, y, θ as before, *what to talk about / how to say it*
- z is a collection of latent variables

Preliminary Example 1: Mixture of Decoders

Generative process:

1. Draw cluster $z \in \{1, \ldots, Z\}$ 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 Example 2: Neural Copy Models

Generative process:

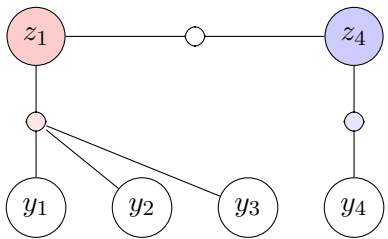
1. Draw copy switch $z \in \{0, 1\}$ from a Bernoulii.
2. Draw words $y_{1:T}$ from decoder RNN where
   - If $z = 0$, let the model generate a new word.
   - If $z = 1$, let the model copy a word from the source.

Example:

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

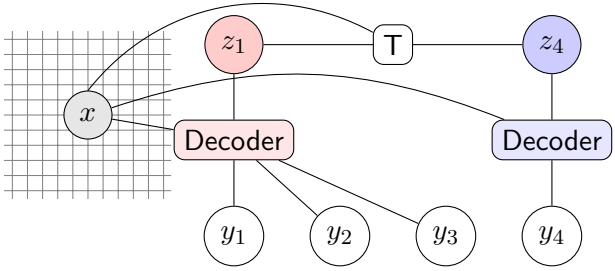
Base Model: Hidden Semi-Markov Model

- 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, $p(y_1, \ldots, y_T, z \mid x)$.
- Transition Distribution: neural network between clusters.
- Emission Distribution: Encoder-Decoder+Copy, specialized per cluster $\{1, \ldots, Z\}$.



Technical Methodology: Training Model

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

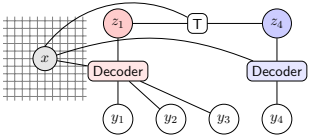
$$\mathcal{L}(\theta) = -\log \sum_{z_{1:T}} p(\hat{y}_{1:T}, z_{1:T} \mid x; \theta)$$

- 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.



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

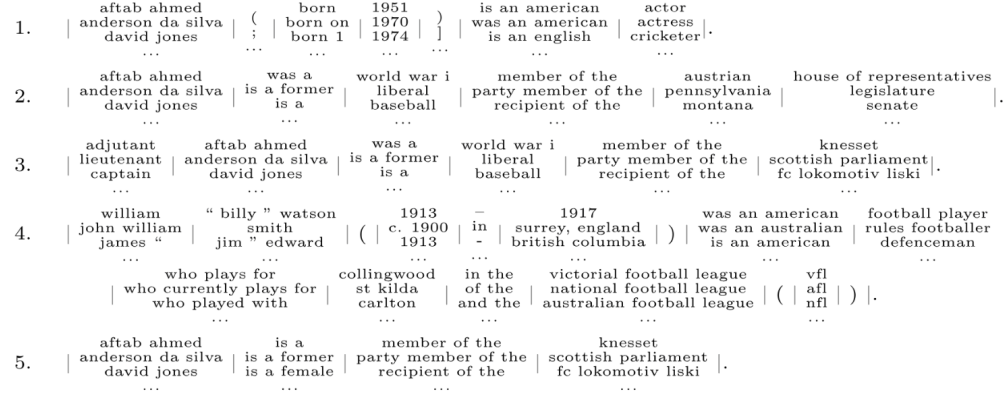
The Wrestlers is a coffee shop that serves English ...

$$\Downarrow z^*$$

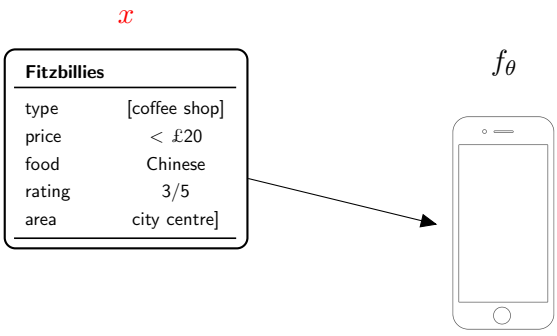
The Wrestlers | is a | coffee shop | that serves | English | ...

Example Templates: Wikipedia

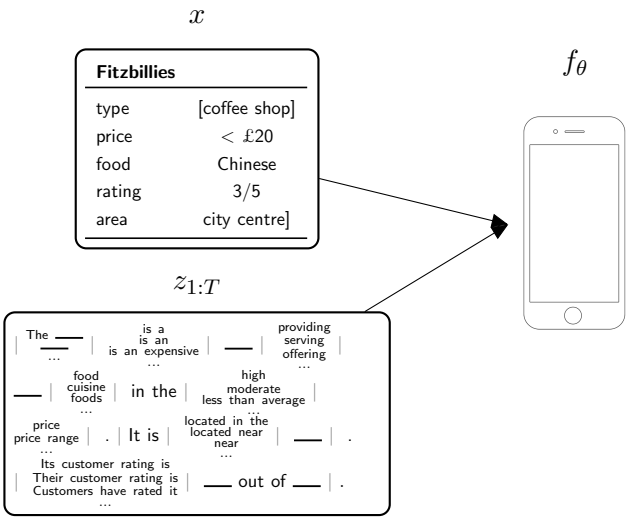
Sentences grouped by the same $z^*_{1:T}$ and their splits.



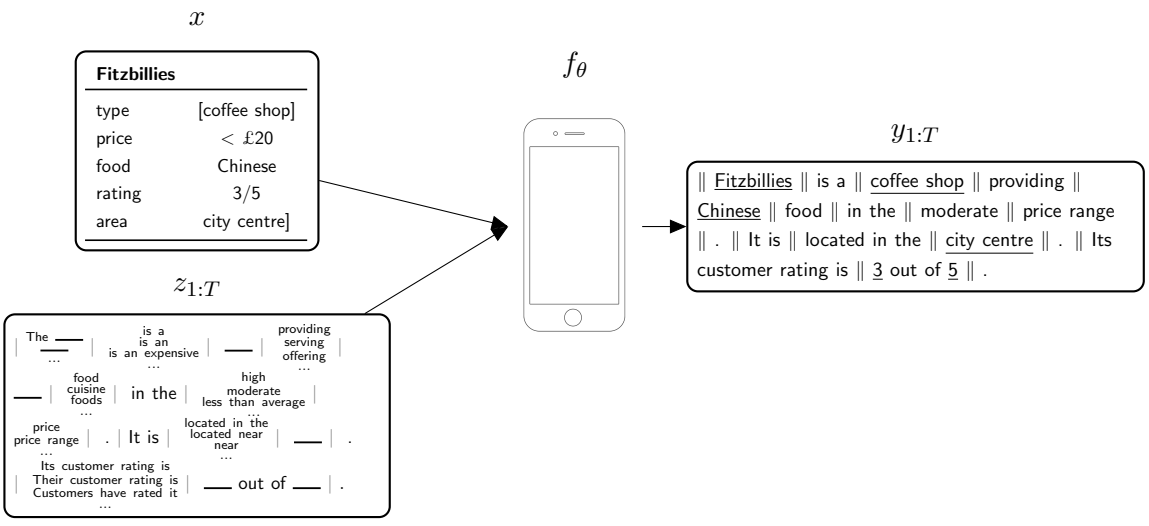
Neural Template Generation Approach



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],

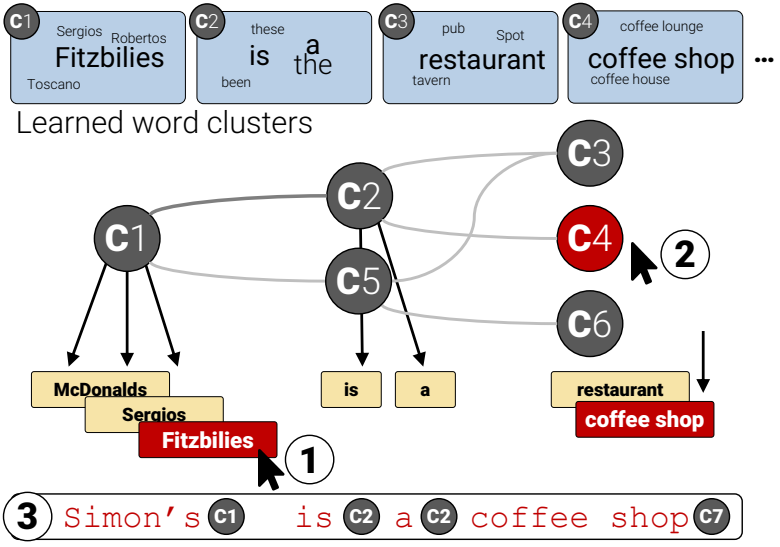
1. The Golden Palace is a coffee shop located in the city centre.
2. In the city centre is a cheap Chinese coffee shop called The Golden Palace.
3. The Golden Palace is a Chinese coffee shop.
4. The Golden Palace is a Chinese coffee shop with a customer rating of 5 out of 5.
5. The Golden Palace that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

Results

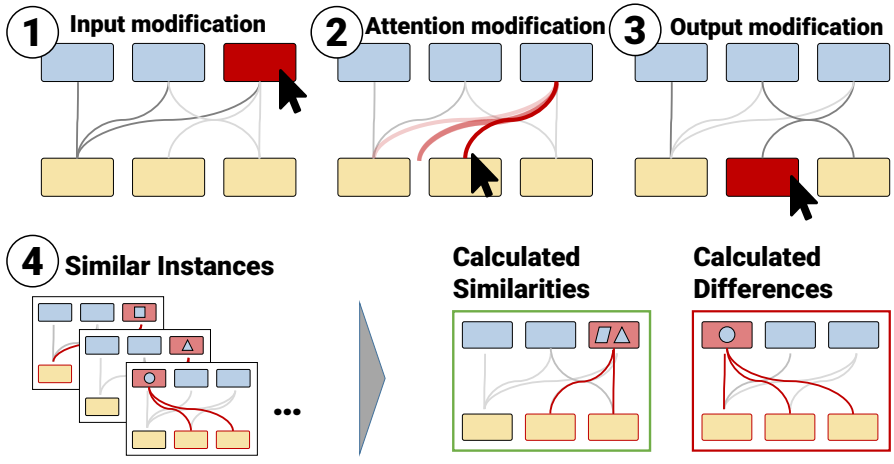
	BLEU	NIST
Test		
Substitution	43.78	6.88
Neural Template	56.72	7.63
Full Neural Model	65.93	8.59

	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

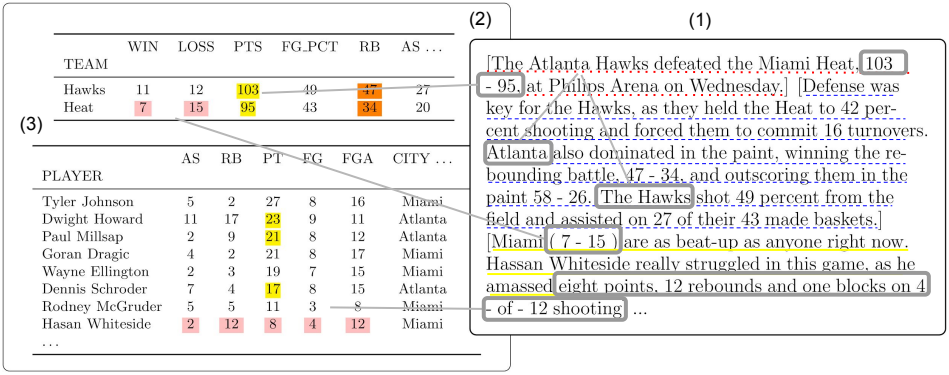
Controllable Interactive Deep Learning Systems



Another Application: Understanding Model Selection



Long-Form Generation with Explicit Reasoning



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

Talk Outline

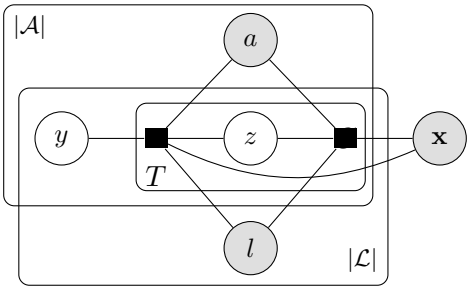
1. Background: Core Model and Implementation
2. Work 1: Rethinking Model Training (*Beam Search Optimization*)
3. Work 2: Rethinking Generation (*Learning Neural Templates*)
4. **Future Challenges Beyond Text Generation**

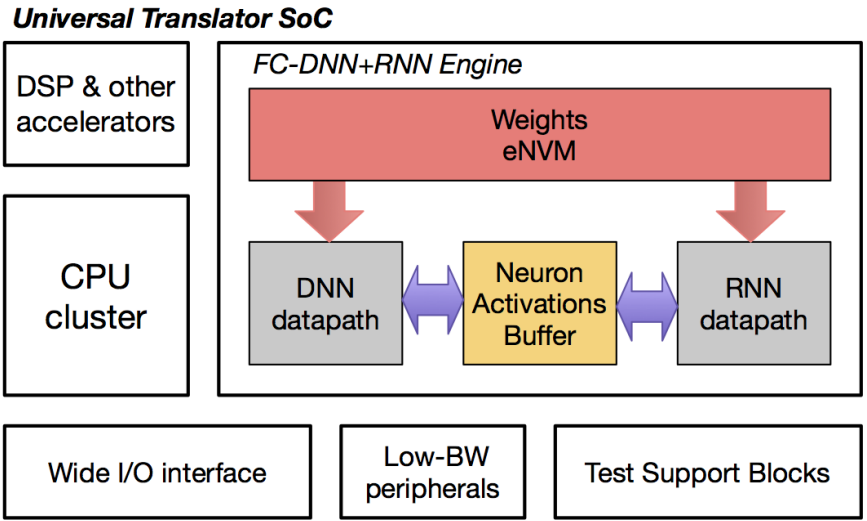
Deep Learning and Natural Language Processing

Simpler and Cleaner Open-Research

Probabilistic Programming

Preprint w/ Uber




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*.