Get To The Point: Summarization with Pointer-Generator Networks (ACL 2017)

Authors: Abigail See, Peter J. Liu, Christopher D. Manning

Presenter: Lu Wang

[Slides modified from paper conference presentation https://www.aclweb.org/anthology/P17-1099/]

Two approaches to summarization

Extractive Summarization

Select parts (typically sentences) of the original text to form a summary.

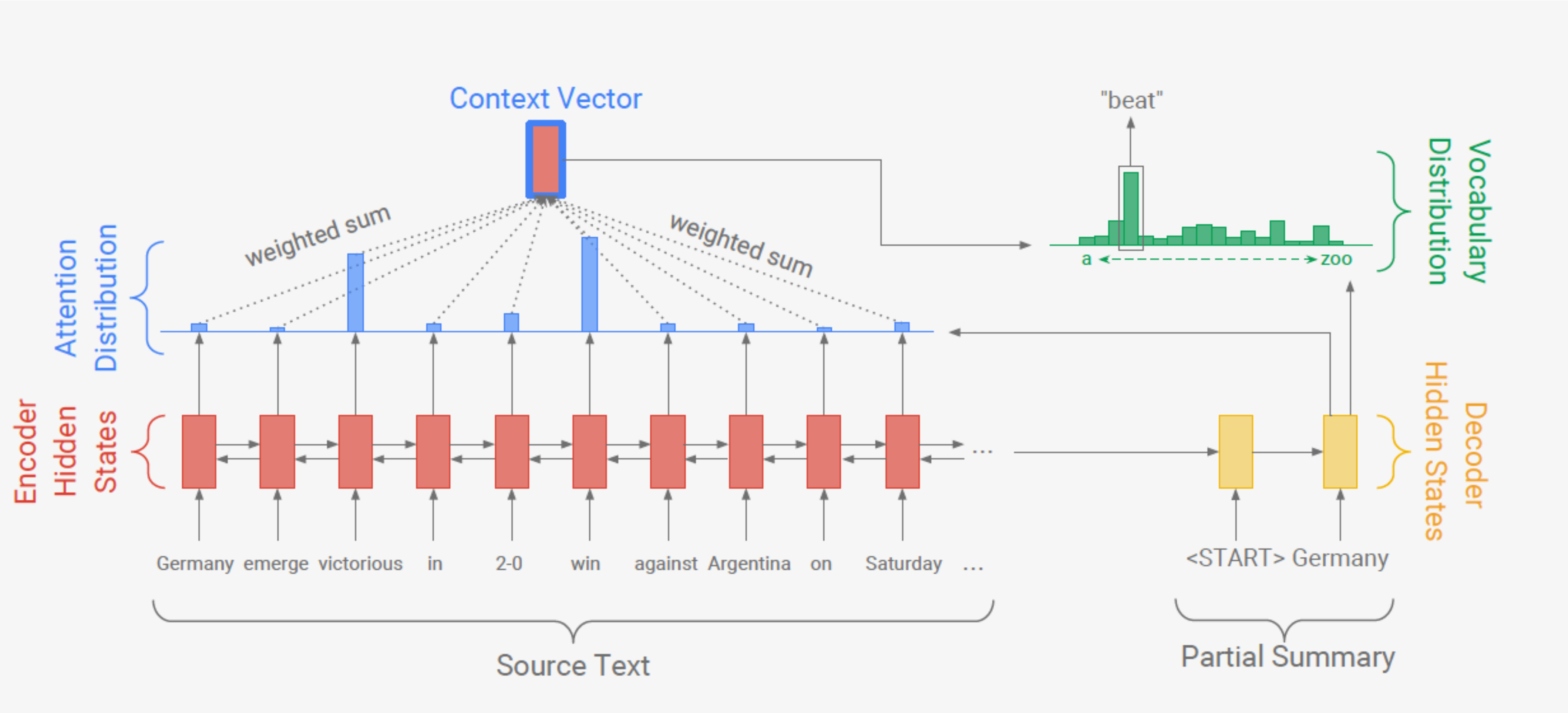
- Easier
- Too restrictive (no paraphrasing)
- Most past work is extractive

Abstractive Summarization

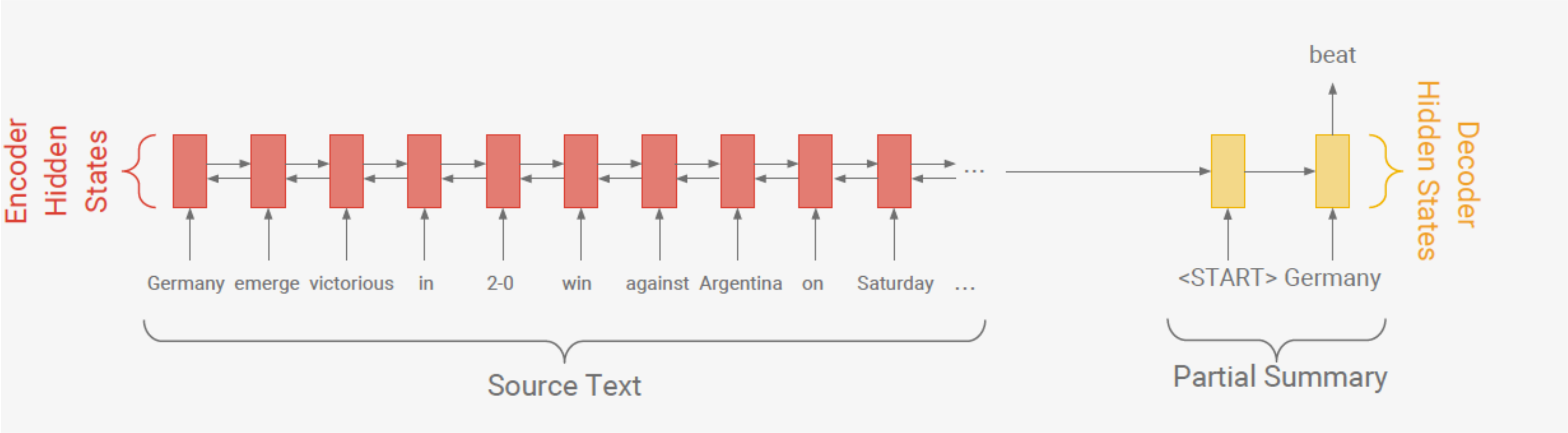
Generate novel sentences using natural language generation techniques.

- More difficult
- More flexible and human
- Necessary for future progress

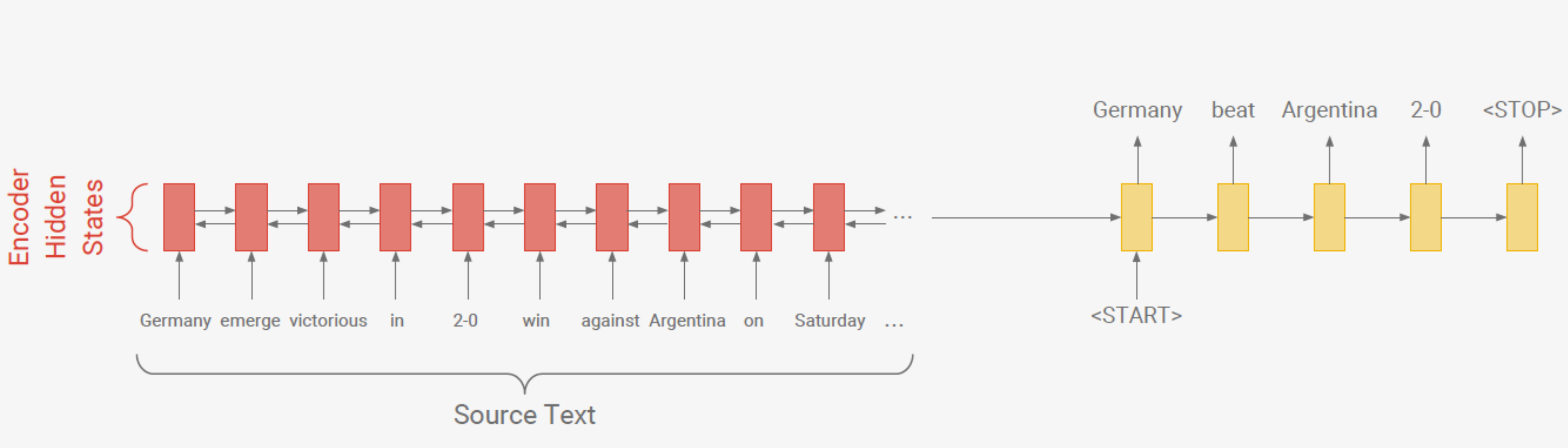
Sequence-to-sequence + attention model



Sequence-to-sequence + attention model



Sequence-to-sequence + attention model



Two problems

Problem 1: The summaries sometimes reproduce factual details inaccurately.

e.g. *Germany beat Argentina 3-2*

Incorrect **rare** or **out-of-vocabulary** word

Problem 2: The summaries sometimes repeat themselves.

e.g. *Germany beat Germany beat Germany beat…*

Two problems

Problem 1: The summaries sometimes reproduce factual details inaccurately.

e.g. *Germany beat Argentina 3-2*

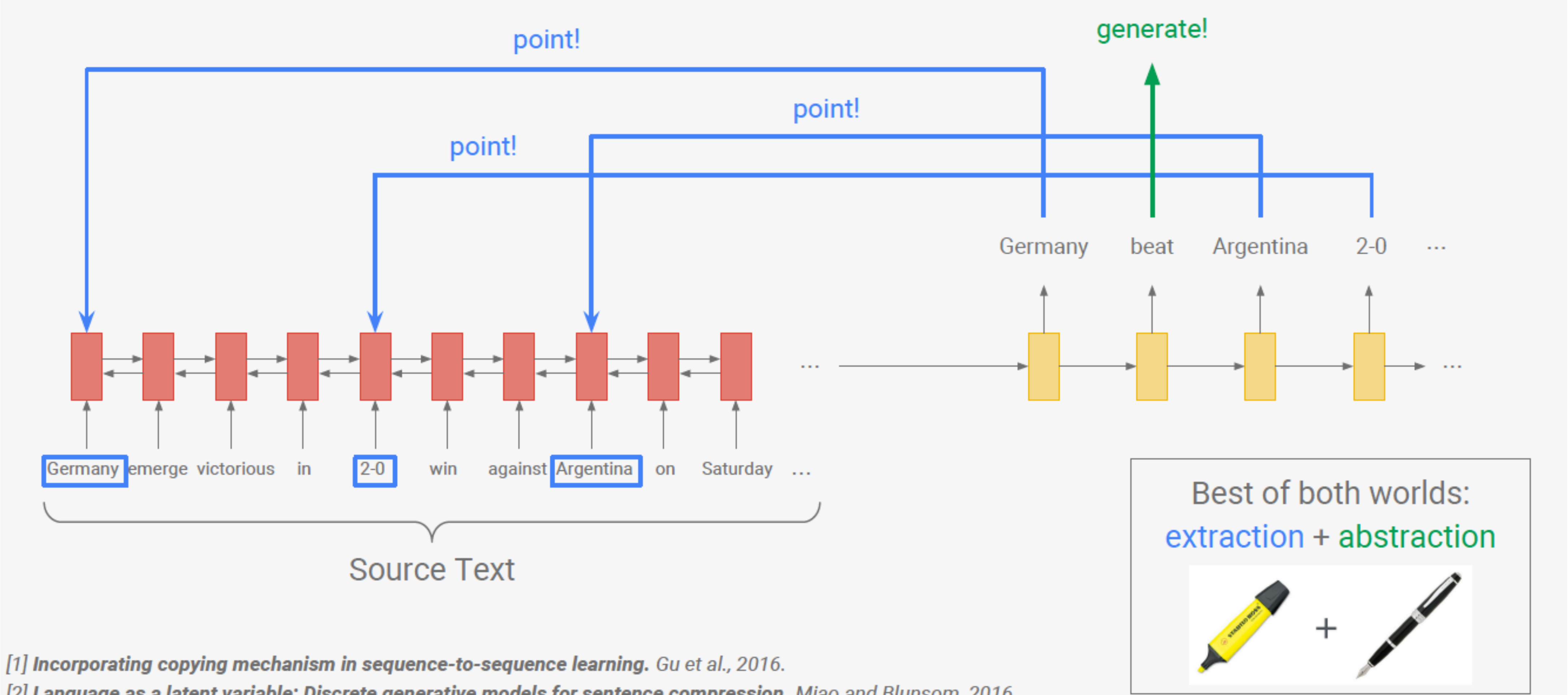
Incorrect **rare** or **out-of-vocabulary** word

Solution: Use a pointer to copy words.

Problem 2: The summaries sometimes repeat themselves.

e.g. *Germany beat Germany beat Germany beat…*

Use pointers



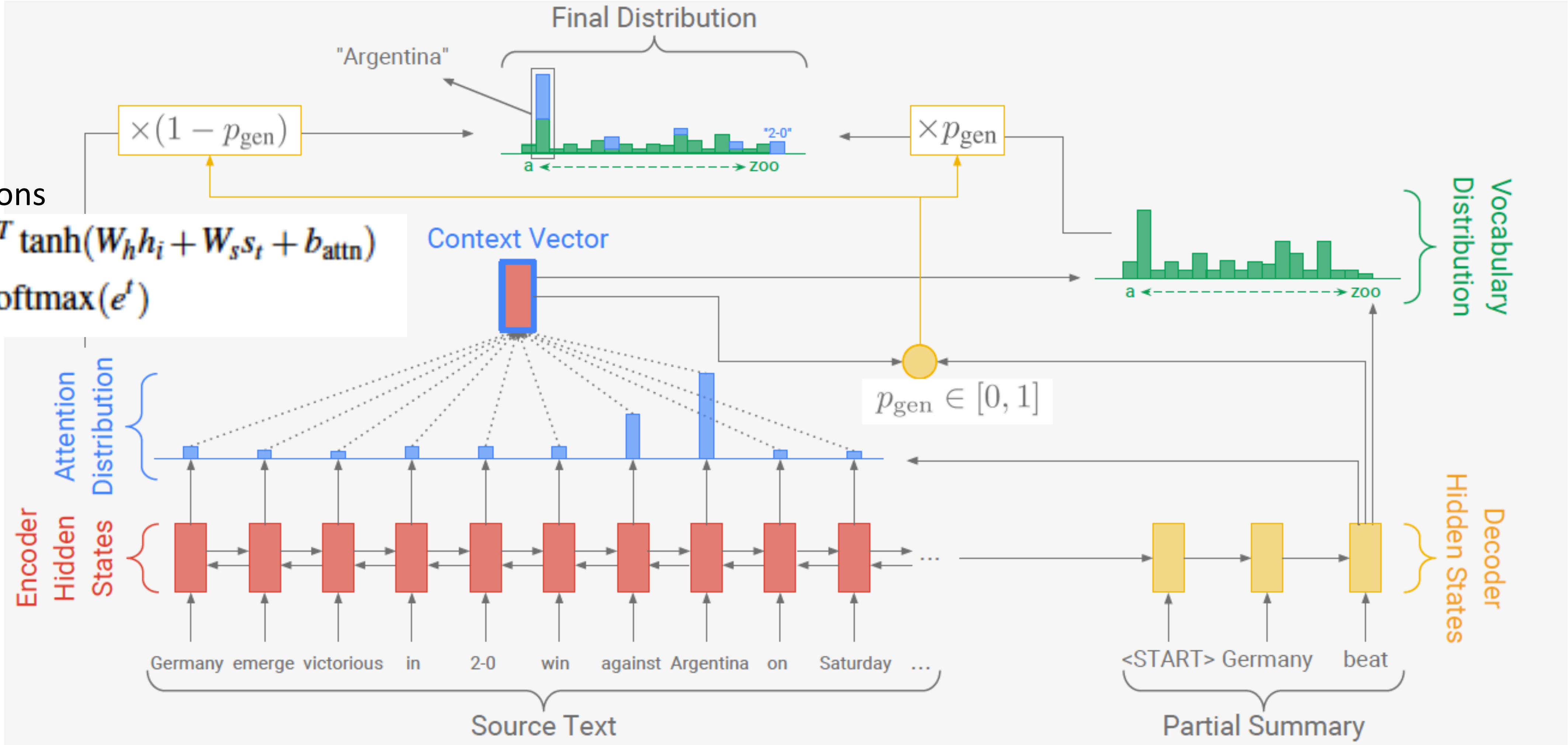
[1] ***Incorporating copying mechanism in sequence-to-sequence learning.*** *Gu et al., 2016.*
[2] ***Language as a latent variable: Discrete generative models for sentence compression.*** *Miao and Blunsom, 2016.*

Pointer-generator network

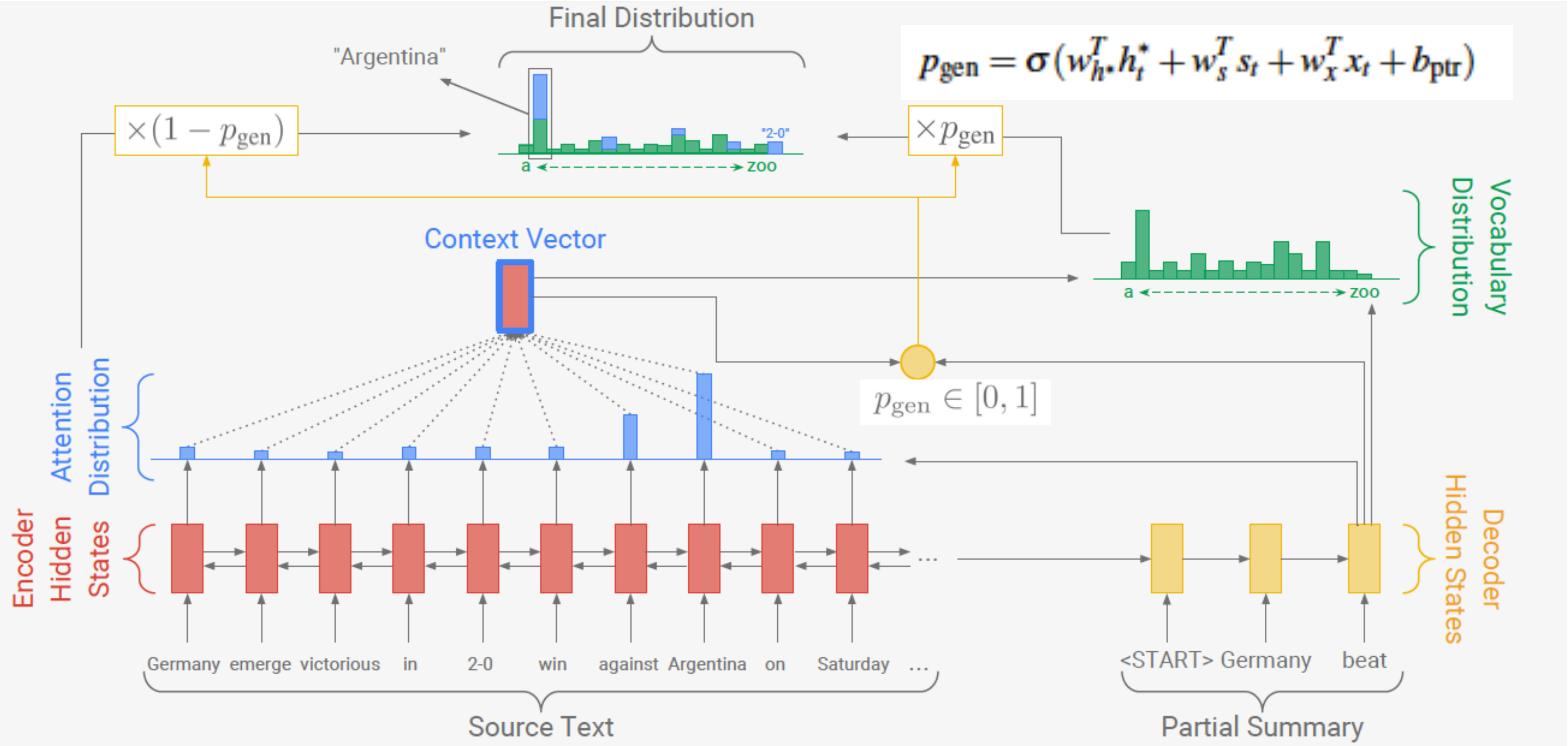
Attentions

$$e_i^t = v^T \tanh(W_h h_i + W_s s_t + b_{\text{attn}})$$

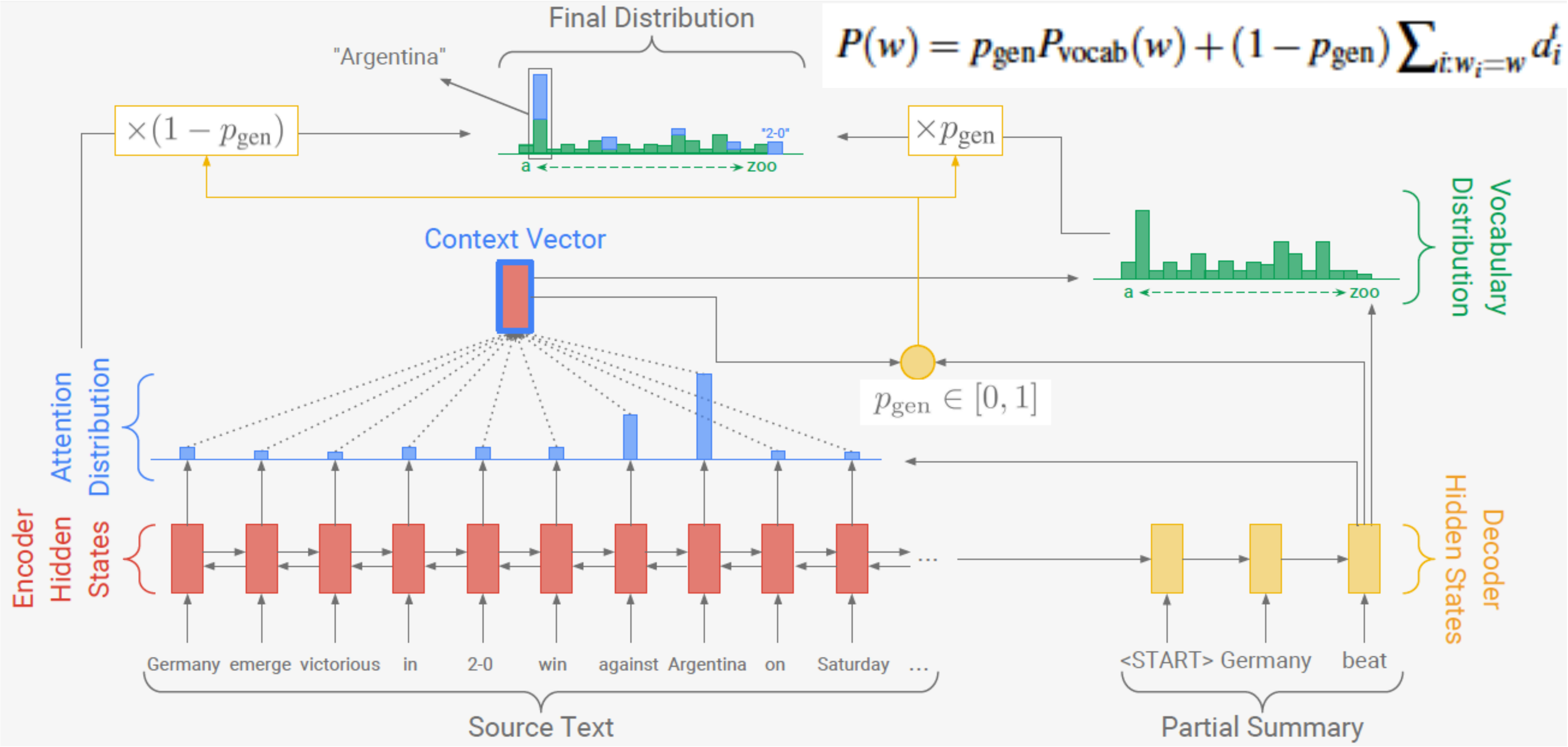
$$a^t = \text{softmax}(e^t)$$



Pointer-generator network



Pointer-generator network



Improvements

Before	After
UNK UNK was expelled from the dubai open chess tournament	*gaioz nigalidze was expelled from the dubai open chess tournament*
the 2015 rio olympic games	*the 2016 rio olympic games*

Two problems

Problem 1: The summaries sometimes reproduce factual details inaccurately.

e.g. *Germany beat Argentina 3-2*

Solution: Use a pointer to copy words.

Problem 2: The summaries sometimes repeat themselves.

e.g. *Germany beat Germany beat Germany beat…*

Solution: Penalize repeatedly attending to same parts of the source text.

Reducing repetition with coverage

Coverage = cumulative attention = what has been covered so far

Source Text:	Germany emerge victorious in 2-0 win against Argentina on Saturday
Summary:	Germany beat _____

[4] ***Modeling coverage for neural machine translation.*** *Tu et al., 2016,*
[5] ***Coverage embedding models for neural machine translation.*** *Mi et al., 2016*
[6] ***Distraction-based neural networks for modeling documents.*** *Chen et al., 2016.*

Reducing repetition with coverage

Coverage = cumulative attention = what has been covered so far

Don't attend here

Source Text: Germany emerge victorious in 2-0 win against Argentina on Saturday

Summary: Germany beat ______

1. Use coverage as extra input to attention mechanism.
2. Penalize attending to things that have already been covered.

[4] ***Modeling coverage for neural machine translation.*** *Tu et al., 2016,*
[5] ***Coverage embedding models for neural machine translation.*** *Mi et al., 2016*
[6] ***Distraction-based neural networks for modeling documents.*** *Chen et al., 2016.*

Reducing repetition with coverage

Coverage = cumulative attention = what has been covered so far

Don't attend here

Source Text: Germany emerge victorious in 2-0 win against Argentina on Saturday

Summary: Germany beat ______

$$c^t = \sum_{t'=0}^{t-1} a^{t'}$$

1. Use coverage as extra input to attention mechanism.
2. Penalize attending to things that have already been covered.

[4] ***Modeling coverage for neural machine translation.*** *Tu et al., 2016,*
[5] ***Coverage embedding models for neural machine translation.*** *Mi et al., 2016*
[6] ***Distraction-based neural networks for modeling documents.*** *Chen et al., 2016.*

Reducing repetition with coverage

Coverage = cumulative attention = what has been covered so far

Don't attend here

Source Text: Germany emerge victorious in 2-0 win against Argentina on Saturday

Summary: Germany beat ______

$$c^t = \sum_{t'=0}^{t-1} a^{t'}$$

1. Use coverage as extra input to attention mechanism.
2. Penalize attending to things that have already been covered.

$$\text{loss}_t = -\log P(w_t^*) + \lambda \sum_i \min(a_i^t, c_i^t) \qquad \text{covloss}_t = \sum_i \min(a_i^t, c_i^t)$$

*[4] **Modeling coverage for neural machine translation.** Tu et al., 2016,*
*[5] **Coverage embedding models for neural machine translation.** Mi et al., 2016*
*[6] **Distraction-based neural networks for modeling documents.** Chen et al., 2016.*

Datasets

- CNN/Daily Mail (Hermann et al., 2015)
  - 287,226 training examples, 13,368 validation examples and 11,490 testing examples
  - limit the input length to 400 tokens and output length to 100 tokens for training and 120 for testing

Results

ROUGE compares the machine-generated summary to the human-written reference summary and counts co-occurrence of 1-grams, 2-grams, and longest common sequence.

	ROUGE-1	ROUGE-2	ROUGE-L
Nallapati et al. 2016	35.5	13.3	32.7
Ours (seq2seq baseline)	31.3	11.8	28.8
Ours (pointer-generator)	36.4	15.7	33.4
Ours (pointer-generator + coverage)	**39.5**	**17.3**	**36.4**

Previous best abstractive result

Our improvements

Results

ROUGE compares the machine-generated summary to the human-written reference summary and counts co-occurrence of 1-grams, 2-grams, and longest common sequence.

	ROUGE-1	ROUGE-2	ROUGE-L
Nallapati et al. 2016	35.5	13.3	32.7
Ours (seq2seq baseline)	31.3	11.8	28.8
Ours (pointer-generator)	36.4	15.7	33.4
Ours (pointer-generator + coverage)	**39.5**	**17.3**	**36.4**
Lead-3 (first three sentences)	40.3	17.7	36.6

Previous best abstractive result

Our improvements

The difficulty of evaluating summarization

- Summarization is **subjective**
  - There are many correct ways to summarize
- ROUGE is based on **strict** comparison to a reference summary
  - Intolerant to rephrasing
  - Rewards extractive strategies
- Take first 3 sentences as summary → higher ROUGE than (almost) any published system
  - Partially due to news article structure

A Deep Reinforced Model for Abstractive Summarization (ICLR 2018)

Authors: Romain Paulus, Caiming Xiong, Richard Socher

Presenter: Lu Wang

[Some figures taken from Paulus' presentation]

Three problems

- Repetitive content in the output (this is discussed in the first paper)

Three problems

- Repetitive content in the output (this is discussed in the first paper)

- Long-term coherence
  - hard to stay on the same topic or show connections when multiple sentences are generated
  - Ordering 1: Lisa went to sail. There was a gale. Lisa came home.
  - Ordering 2: Lisa came home. There was a gale. Lisa went to sail.

Three problems

- Repetitive content in the output (this is discussed in the first paper)

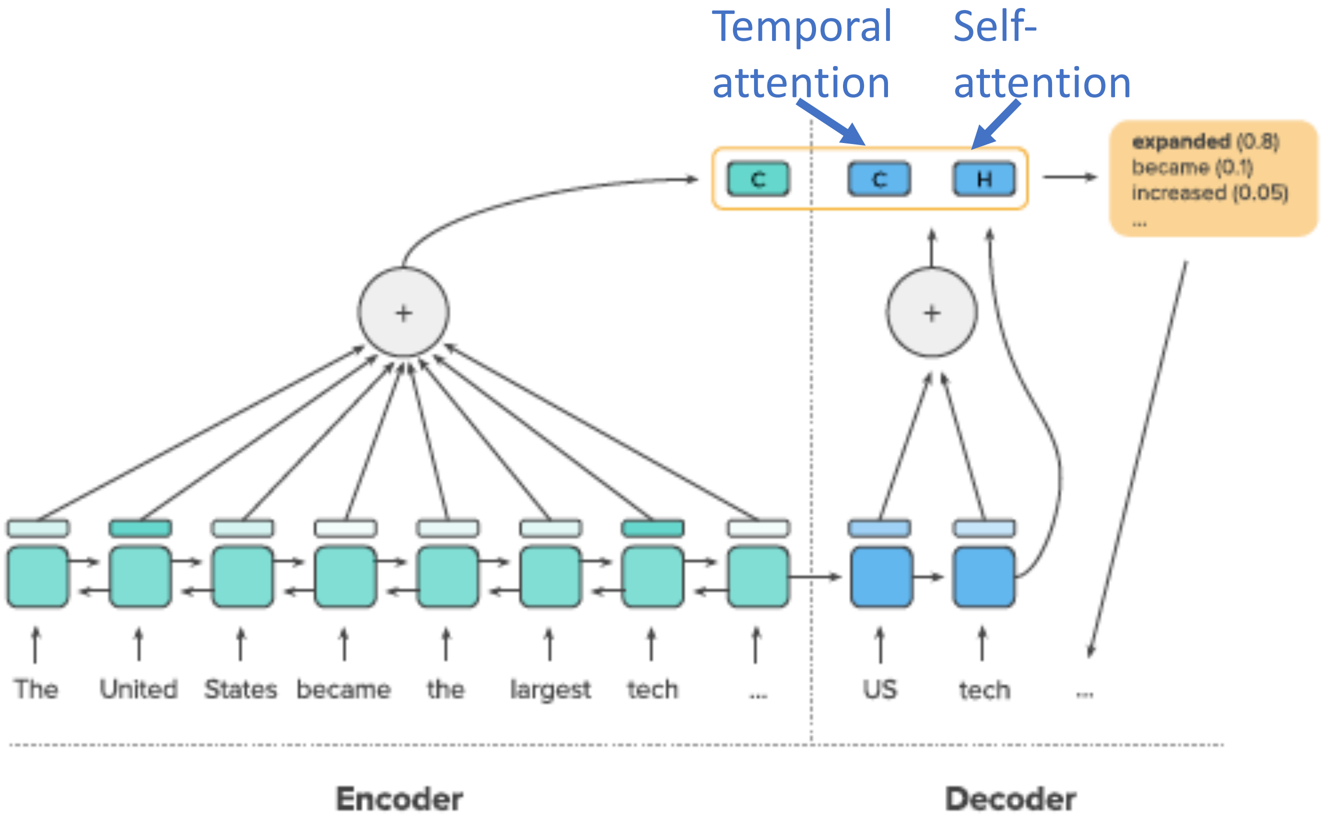
- Long-term coherence
  - hard to stay on the same topic or show connections when multiple sentences are generated
  - Ordering 1: Lisa went to sail. There was a gale. Lisa came home.
  - Ordering 2: Lisa came home. There was a gale. Lisa went to sail.

- Directly optimize on ROUGE scores
  - ROUGE measure word overlaps between system generated summaries and human-written summaries
  - existing training objective use likelihood of each generated token, i.e. $p(y_t|x)$

Three problems

- Repetitive content in the output
- Long-term coherence
- Directly optimize on ROUGE scores

Temporal attention on the input + self-attention on the output



Temporal attention

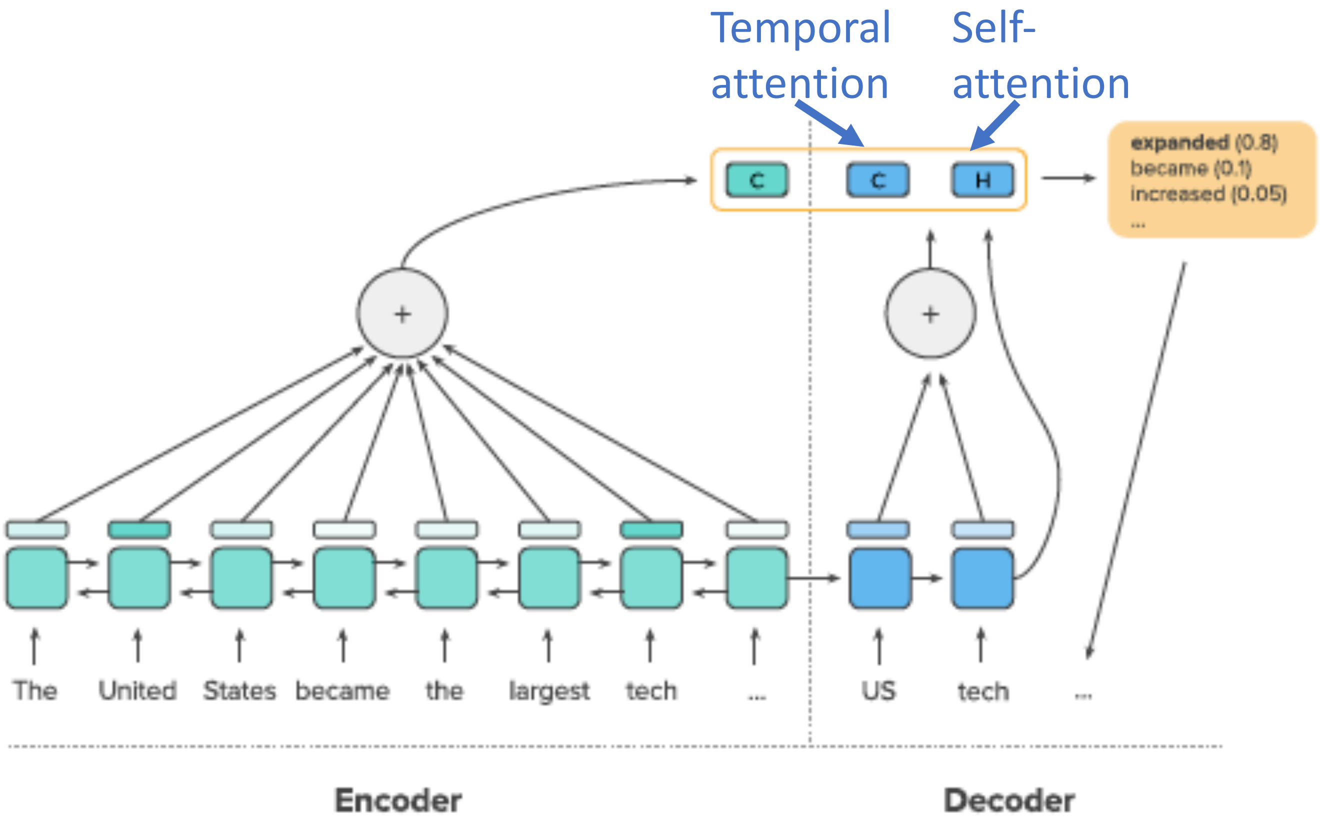
- Input attention weights (different from pointer-generator paper)

$$e_{ti} = f(h_t^d, h_i^e)$$

$$f(h_t^d, h_i^e) = {h_t^d}^T W_{\text{attn}}^e h_i^e$$

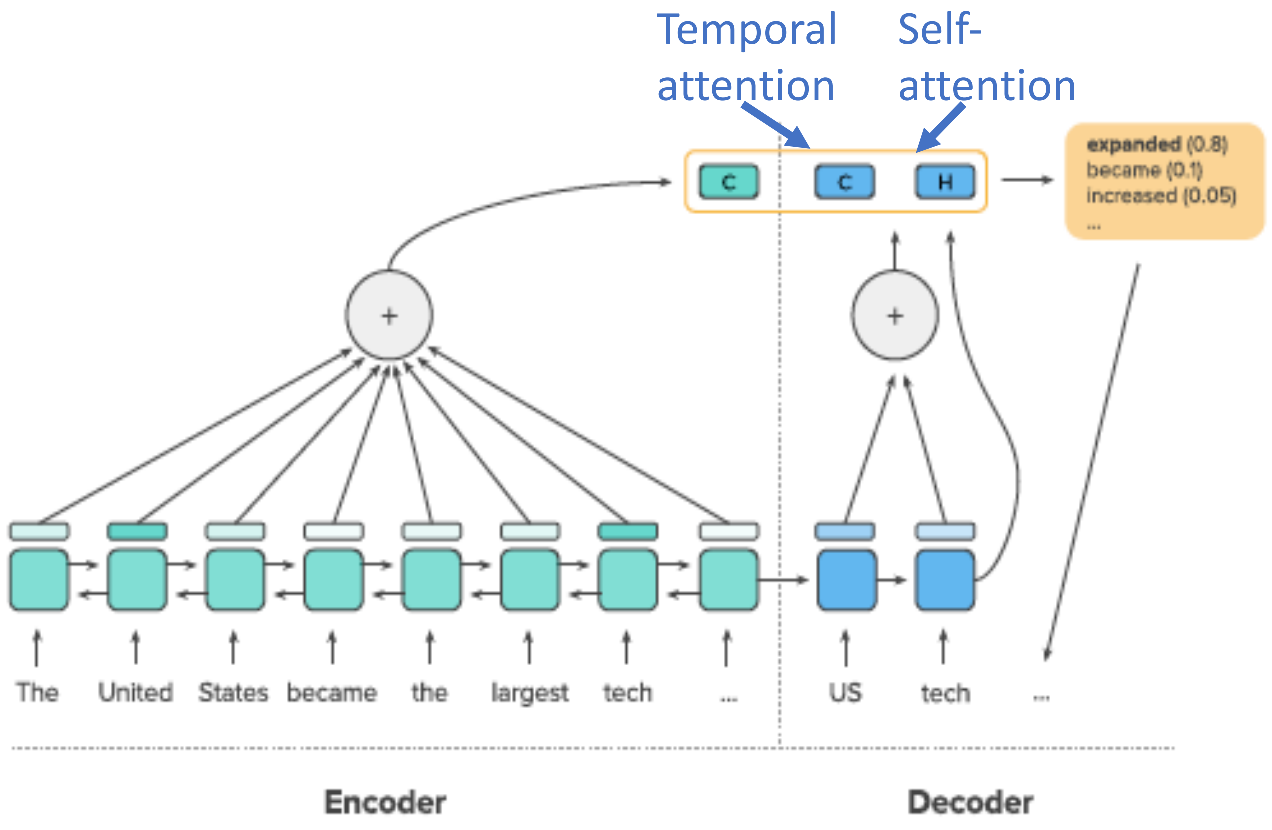
- Penalizing the tokens that obtained high attentions

$$e'_{ti} = \begin{cases} exp(e_{ti}) & \text{if } t = 1 \\ \frac{exp(e_{ti})}{\sum_{j=1}^{t-1} \exp(e_{ji})} & \text{otherwise} \end{cases}$$



Self-attention (intra-decoder attention)

- How to be aware of what has been generated?



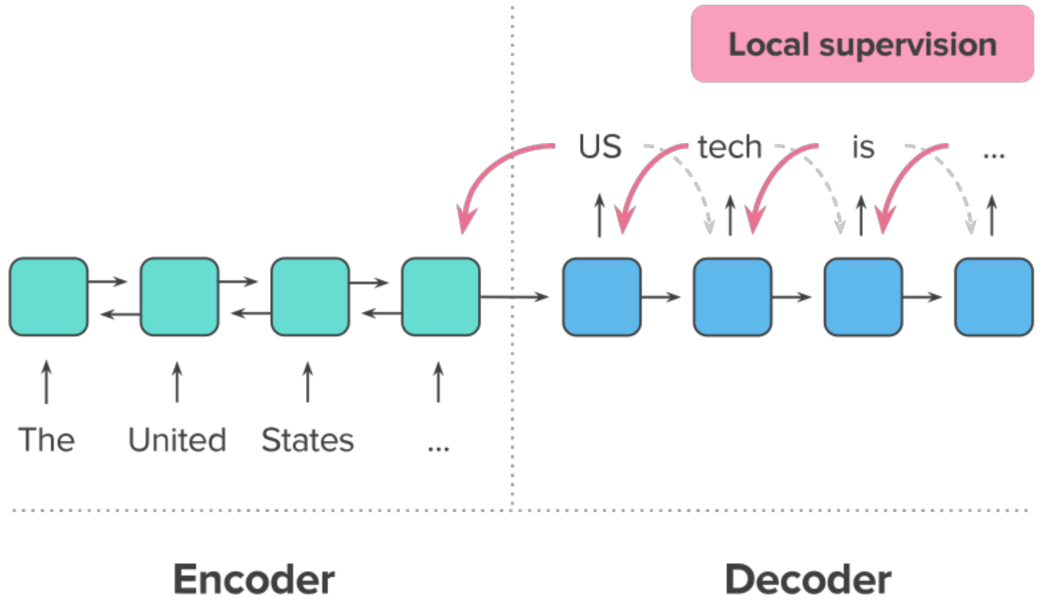
Self-attention (intra-decoder attention)

$$c_t^d = \sum_{j=1}^{t-1} \alpha_{tj}^d h_j^d$$

Consider what has been generated

$$e_{tt'}^d = {h_t^d}^T W_{\text{attn}}^d h_{t'}^d$$

$$\alpha_{tt'}^d = \frac{exp(e_{tt'}^d)}{\sum_{j=1}^{t-1} exp(e_{tj}^d)}$$

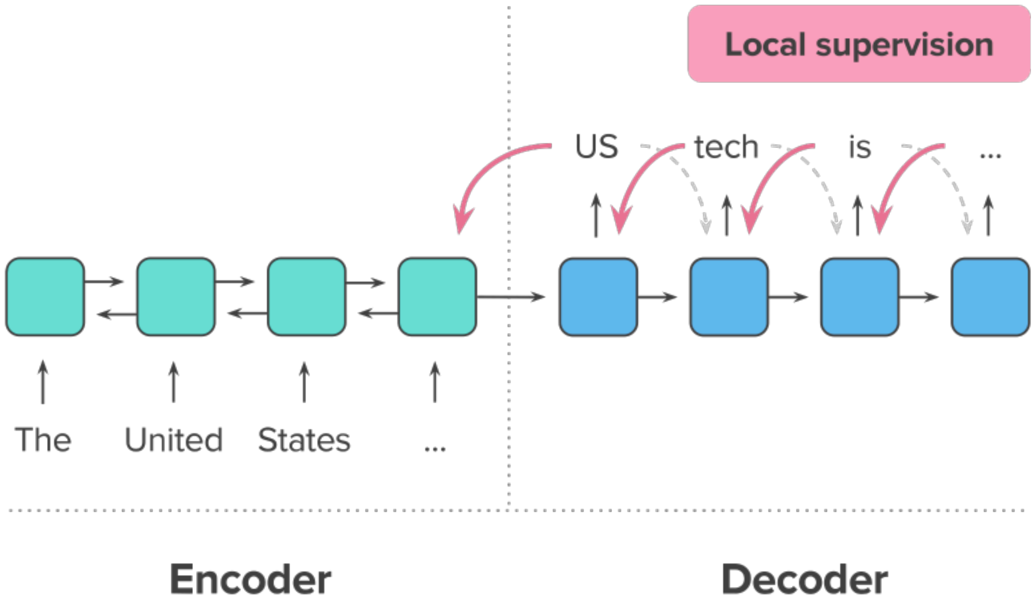


Self-attention (intra-decoder attention)

$$e^d_{tt'} = {h^d_t}^T W^d_{\text{attn}} h^d_{t'}$$

$$\alpha^d_{tt'} = \frac{exp(e^d_{tt'})}{\sum_{j=1}^{t-1} exp(e^d_{tj})}$$

$$c^d_t = \sum_{j=1}^{t-1} \alpha^d_{tj} h^d_j$$



$$p(y_t|u_t = 0) = \text{softmax}(W_{\text{out}}[h^d_t \| c^e_t \| c^d_t] + b_{\text{out}})$$

Input attention   self-attention

Three problems

- Repetitive content in the output
- Long-term coherence
- Directly optimize on ROUGE scores

Global reward with ROUGE

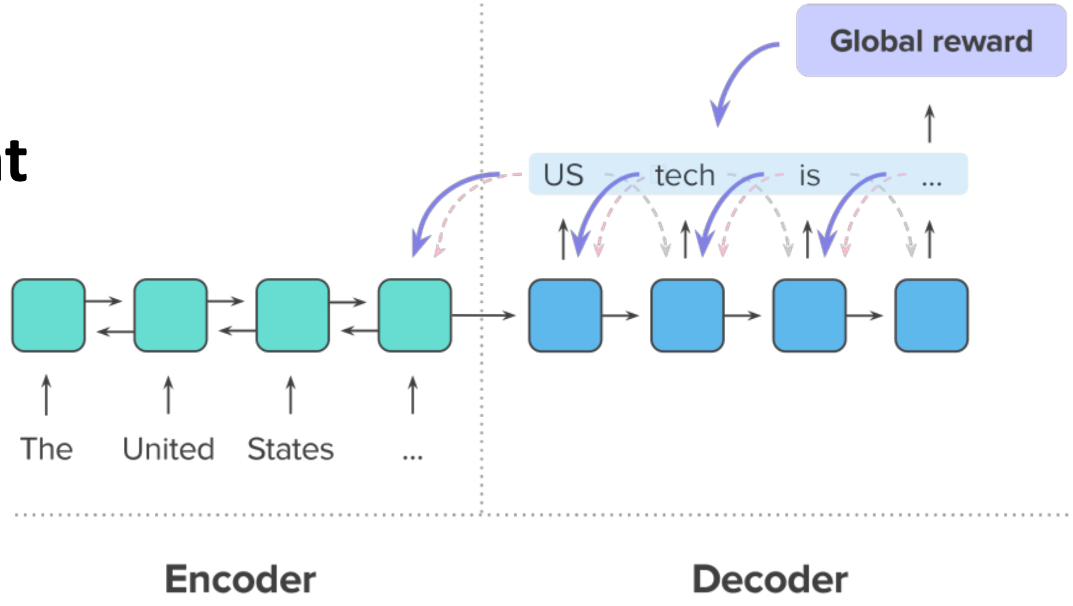
- Idea: directly using ROUGE scores as reward
- But ROUGE is not differentiable

- Training method: self-critical **policy gradient** training algorithm

$$L_{rl} = (r(\hat{y}) - r(y^s)) \sum_{t=1}^{n'} \log p(y_t^s | y_1^s, \ldots, y_{t-1}^s, x)$$

Baseline: greedy decoding

Sampled sequence



- Intuitively, we aim to maximize the conditional likelihood of the **sampled sequence** y^s if it obtains a higher reward than the **baseline**

New training objective

$$L_{mixed} = \gamma L_{rl} + (1 - \gamma) L_{ml}$$

New reinforcement learning objective

Regular log-likelihood objective

Datasets for experiments

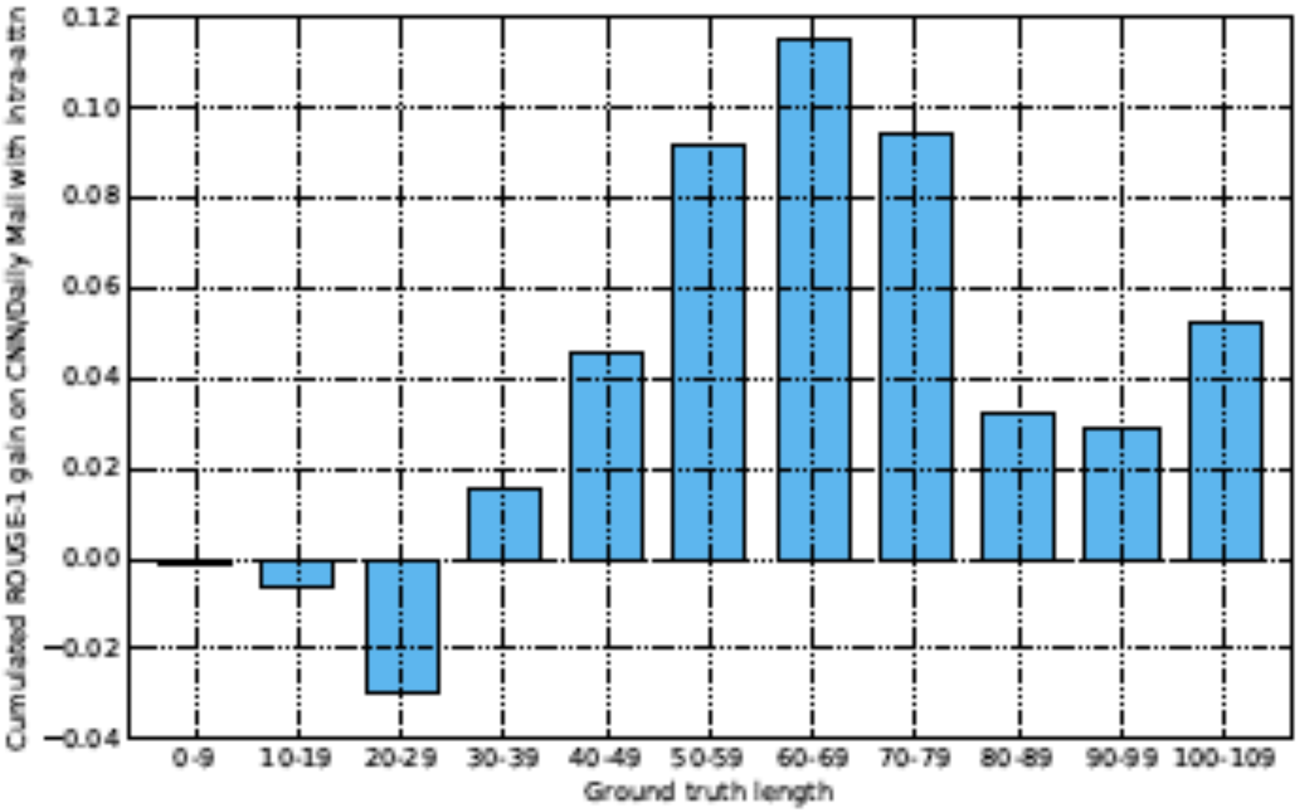
- CNN/Daily Mail (Hermann et al., 2015)
  - 287,113 training examples, 13,368 validation examples and 11,490 testing examples
  - limit the input length to 800 tokens and output length to 100 tokens

- New York Times (Sandhaus, 2008)
  - 589,284 examples for training, 32,736 for validation, and 32,739 for testing

Results on CNN/Daily Mail

Model	ROUGE-1	ROUGE-2	ROUGE-L
Lead-3 (Nallapati et al., 2017)	39.2	15.7	35.5
SummaRuNNer (Nallapati et al., 2017)	39.6	16.2	35.3
words-lvt2k-temp-att (Nallapati et al., 2016)	35.46	13.30	32.65
ML, no intra-attention	37.86	14.69	34.99
ML, with intra-attention	38.30	14.81	35.49
RL, with intra-attention	**41.16**	15.75	**39.08**
ML+RL, with intra-attention	39.87	**15.82**	36.90

ROUGE-1 improvement by adding intra-attention on CNN/Daily Mail



Results on NYT

Model	ROUGE-1	ROUGE-2	ROUGE-L
ML, no intra-attention	44.26	27.43	40.41
ML, with intra-attention	43.86	27.10	40.11
RL, no intra-attention	**47.22**	30.51	**43.27**
ML+RL, no intra-attention	47.03	**30.72**	43.10

Sample outputs

Ground truth summary
Button denied 100th race start for McLaren after ERS failure. Button then spent much of the Bahrain Grand Prix on Twitter delivering his verdict on the action as it unfolded. Lewis Hamilton has out-qualified and finished ahead of Mercedes team-mate Nico Rosberg at every race this season. Bernie Ecclestone confirms F1 will make its bow in Azerbaijan next season.
ML, with intra-attention (ROUGE-1 41.58)
Button was denied his 100th race for McLaren. ERS prevented him from making it to the start-line. The Briton. He quizzed after Nico Rosberg accused Lewis Hamilton of pulling off such a manoeuvre in China. Button has been in Azerbaijan for the first time since 2013.
RL, with intra-attention (ROUGE-1 50.00)
Button was denied his 100th race for McLaren after an ERS prevented him from making it to the start-line. It capped a miserable weekend for the Briton. Button has out-qualified. Finished ahead of Nico Rosberg at Bahrain. Lewis Hamilton has. In 11 races. . The race. To lead 2,000 laps. . In. . . And. .
ML+RL, with intra-attention (ROUGE-1 44.00)
Button was denied his 100th race for McLaren. The ERS prevented him from making it to the start-line. Button was his team mate in the 11 races in Bahrain. He quizzed after Nico Rosberg accused Lewis Hamilton of pulling off such a manoeuvre in China.

Human evaluation on CNN/Daily Mail

Model	Readability	Relevance
ML	6.76	7.14
RL	4.18	6.32
ML+RL	**7.04**	**7.45**

Conclusion

- Intra-decoder attention helps with long summary generation
- Reinforcement learning with ROUGE as reward improves performance
- Simply using reinforcement learning hurts readability