Get To The Point: Summarization with Pointer-Generator Networks

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

Presenter：Wenxin Hou

2020.08.03

- Sequence-to-sequence models have made abstractive summarization viable by applying recurrent neural networks (RNNs) to read and freely generate text

- However, current systems exhibit undesirable behavior such as producing **factual errors** and the problem of **repetition**

- **Out-of-vocabulary (OOV)** words also become a problem for existing models

Original Text (truncated): lagos, nigeria (cnn) a day after winning nigeria's presidency, ***muhammadu buhari*** told cnn's christiane amanpour that **he plans to aggressively fight corruption that has long plagued nigeria** and go after the root of the nation's unrest. *buhari* said he'll "rapidly give attention" to curbing violence in the northeast part of nigeria, where the terrorist group boko haram operates. by cooperating with neighboring nations chad, cameroon and niger, **he said his administration is confident it will be able to thwart criminals** and others contributing to nigeria's instability. for the first time in nigeria's history, the opposition defeated the ruling party in democratic elections. *buhari* defeated incumbent goodluck jonathan by about 2 million votes, according to nigeria's independent national electoral commission. **the win comes after a long history of military rule, coups and botched attempts at democracy in africa's most populous nation.**

Baseline Seq2Seq + Attention: **UNK UNK** says his administration is confident it will be able to **destabilize nigeria's economy**. **UNK** says his administration is confident it will be able to thwart criminals and other **nigerians**. **he says the country has long nigeria and nigeria's economy.**

Figure 1: An example article and summarization generated by baseline model

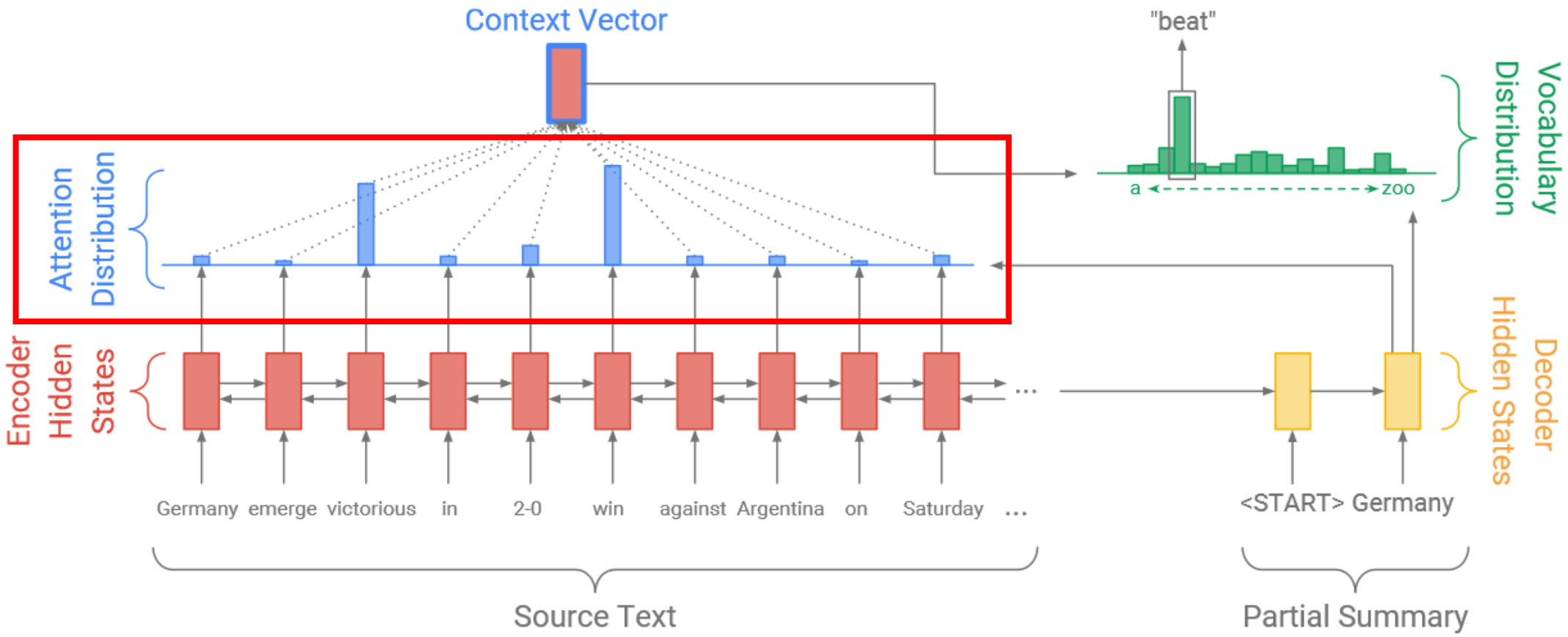
- To handle OOV words and improve accuracy, copying mechanism is introduced to copying words from the source text

- Coverage mechanism is proposed to eliminate repetition

Pointer-Gen: *muhammadu buhari* says he plans to aggressively fight corruption **in the northeast part of nigeria**. he says he'll "rapidly give attention" to curbing violence **in the northeast part of nigeria**. he says his administration is confident it will be able to thwart criminals.

Pointer-Gen + Coverage: *muhammadu buhari* says he plans to aggressively fight corruption that has long plagued nigeria. he says his administration is confident it will be able to thwart criminals. the win comes after a long history of military rule, coups and botched attempts at democracy in africa's most populous nation.

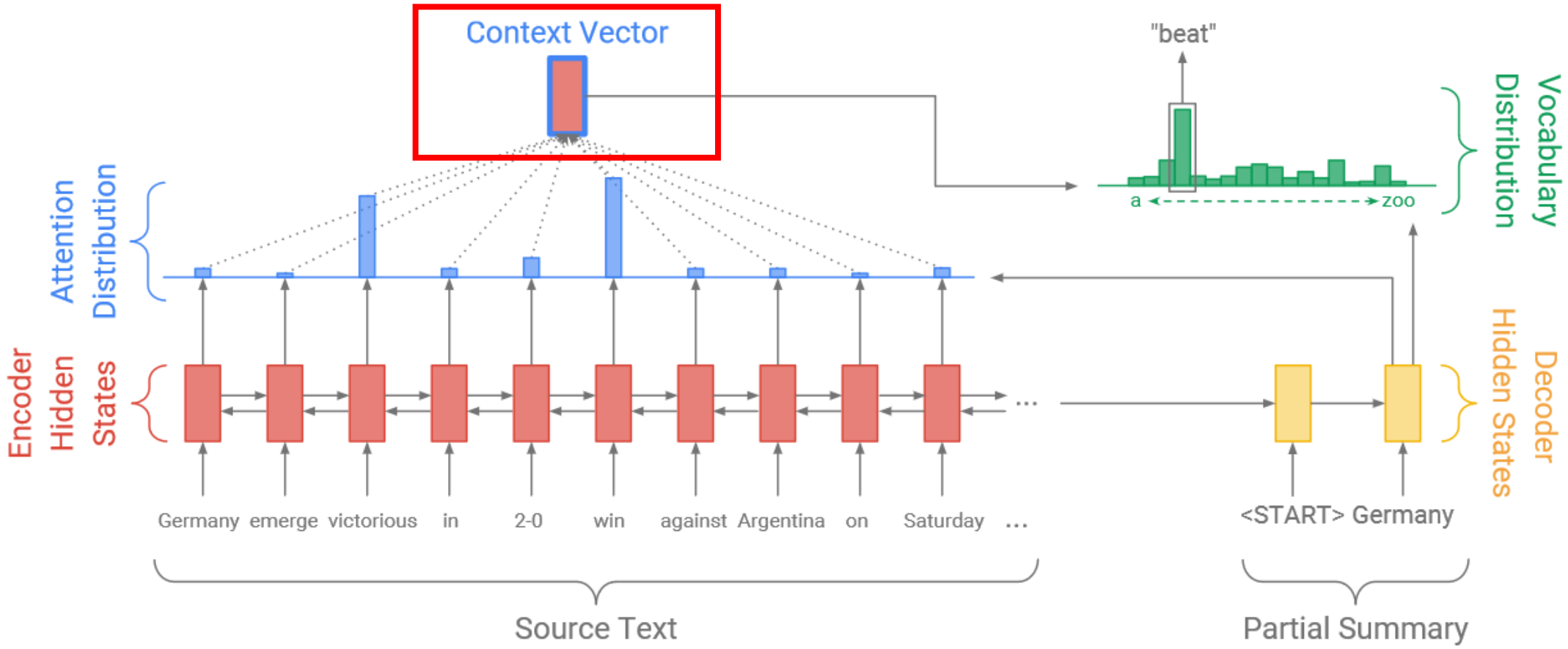
Figure 2: Summarization of example in Figure 1 generated by the proposed models



- **Attention distribution**: a^t at decoder timestep t over the encoder hidden states h_i

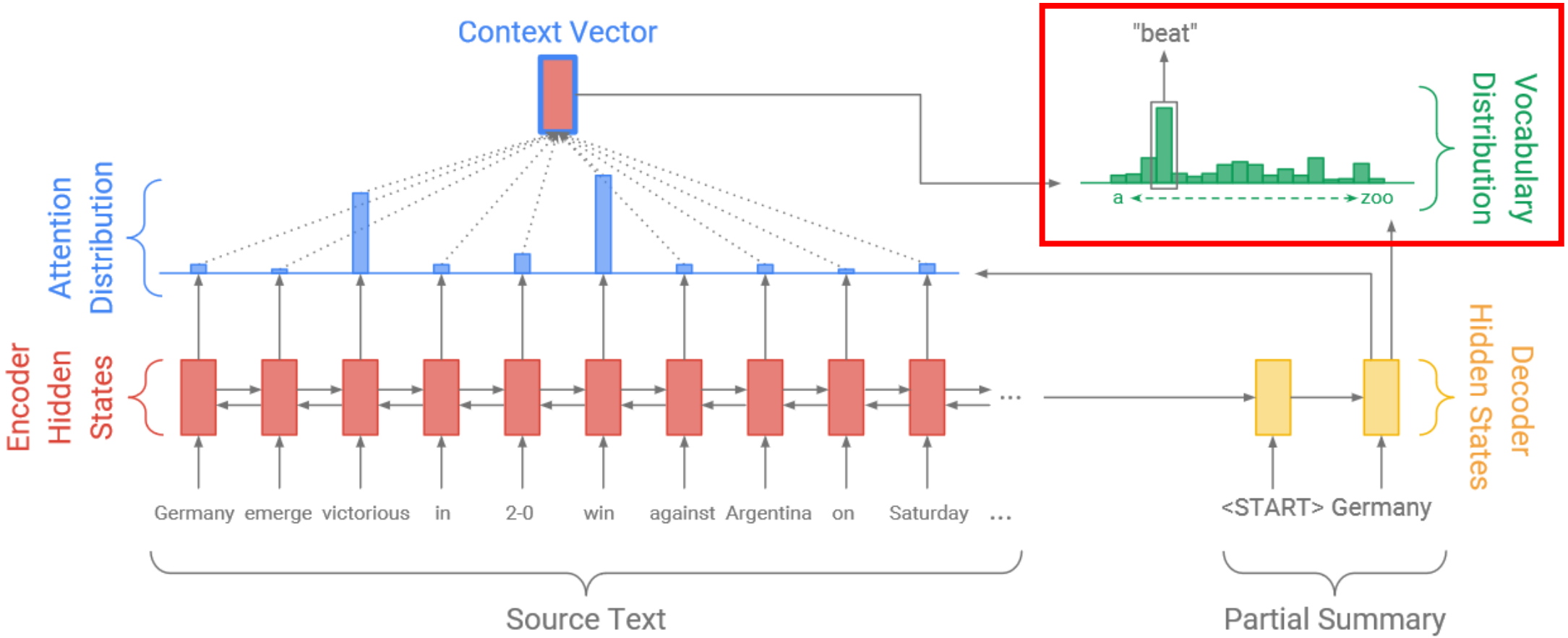
$$e_i^t = v^T \tanh(W_h h_i + W_s s_t + b_{\text{attn}}) \qquad (1)$$

$$a^t = \text{softmax}(e^t) \qquad (2)$$



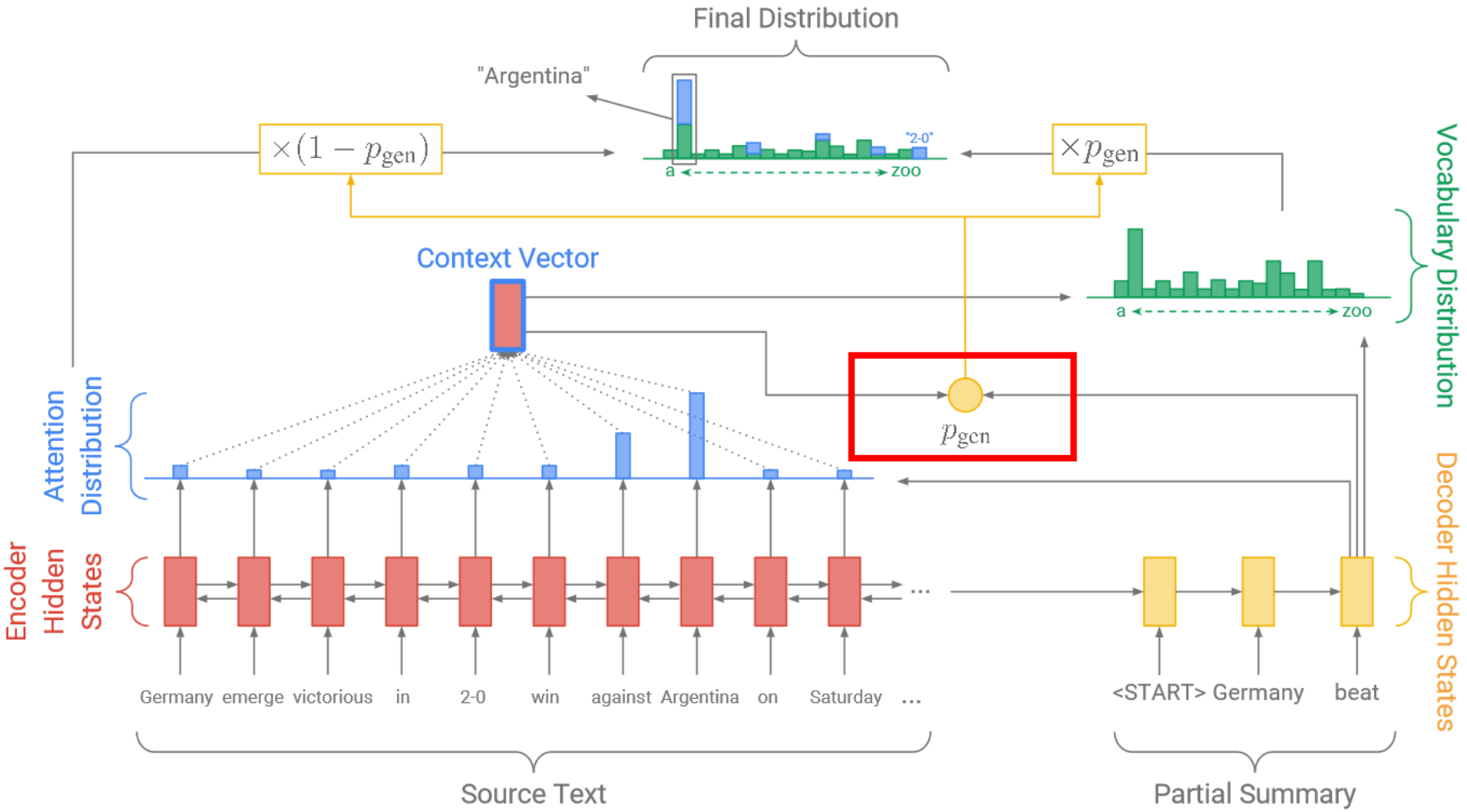
- **Context Vector:** weighted average of encoder hidden states h_i with attention distribution a_t

$$h_t^* = \sum_i a_i^t h_i \qquad (3)$$



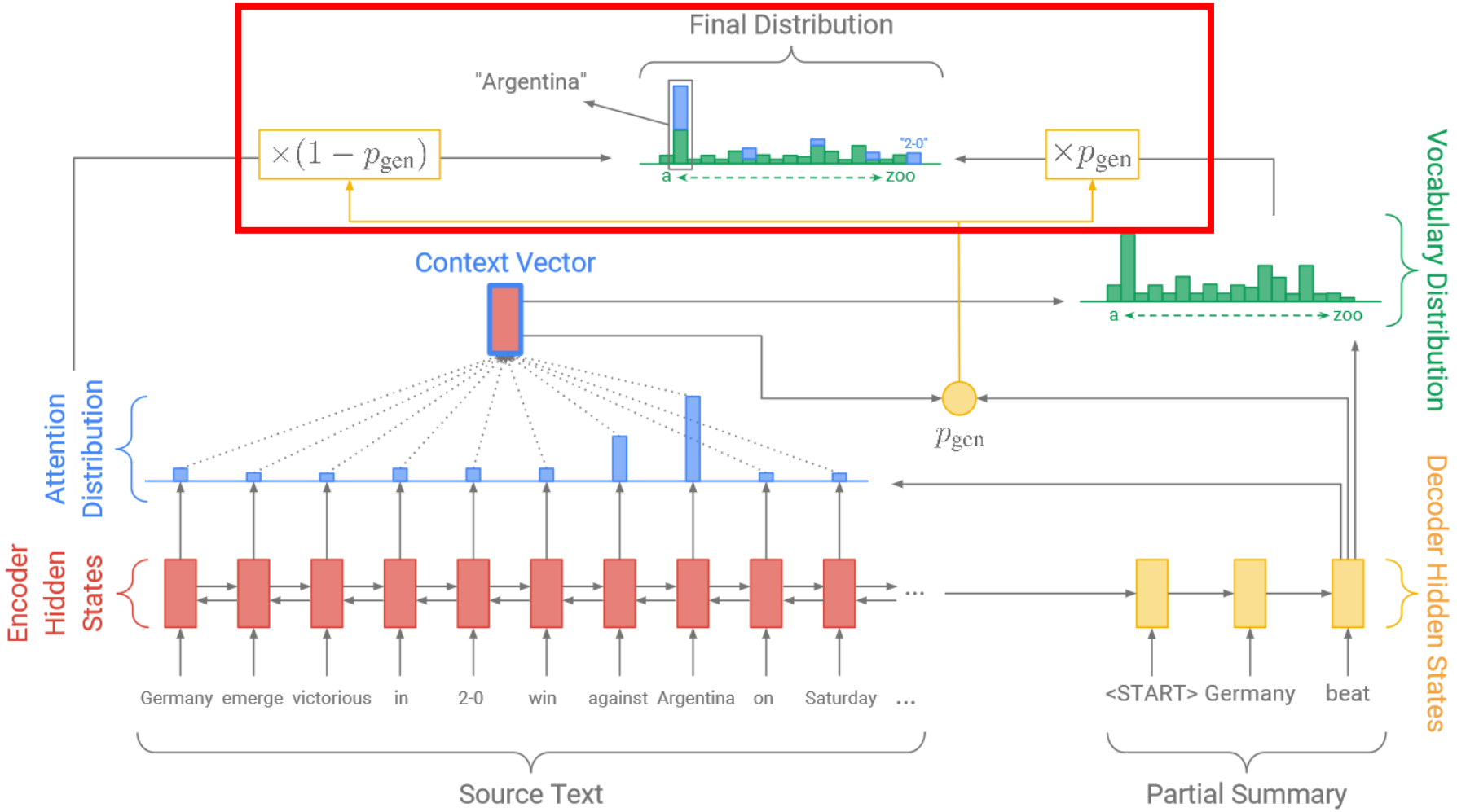
- **Vocabulary distribution:** probability distribution over all words in the vocabulary

$$P_{\text{vocab}} = \text{softmax}(V'(V[s_t, h_t^*] + b) + b') \qquad (4)$$



- **Generation probability:** probability of generating a word from the vocabulary

$$p_{\text{gen}} = \sigma(w_{h^*}^T h_t^* + w_s^T s_t + w_x^T x_t + b_{\text{ptr}}) \qquad (8)$$



- **Final distribution:** probability distribution over the extended vocabulary

$$P(w) = p_{gen} P_{vocab}(w) + (1 - p_{gen}) \sum_{i:w_i=w} a_i^t \quad (9)$$

- **Coverage vector:** sum of attention distributions over all previous decoder timesteps

$$c^t = \sum_{t'=0}^{t-1} a^{t'} \qquad (10)$$

- Coverage vector is used as an extra input to the attention mechanism:

$$e_i^t = v^T \tanh(W_h h_i + W_s s_t + w_c c_i^t + b_{\text{attn}}) \qquad (11)$$

- Coverage loss is added to the primary loss function to penalize the overlap between each attention distribution and the coverage so far :

$$\text{loss}_t = -\log P(w_t^*) + \lambda \sum_i \min(a_i^t, c_i^t) \qquad (13)$$

$$\text{loss} = \frac{1}{T} \sum_{t=0}^{T} \text{loss}_t \qquad (7)$$

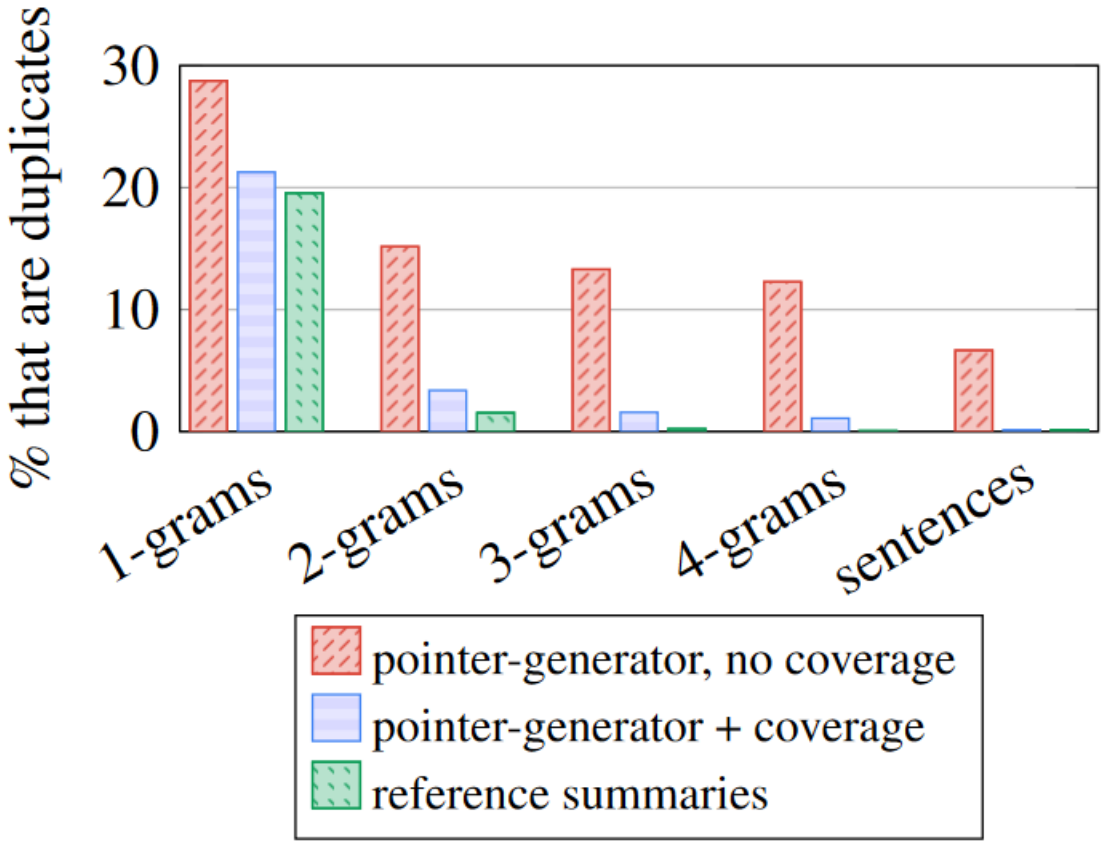
- **Dataset:** CNN/Daily Mail (287,226 training pairs, 13,368 validation pairs and 11,490 test pairs)
- **Evaluation metrics:** ROUGE [1], METEOR [2] (higher is better)

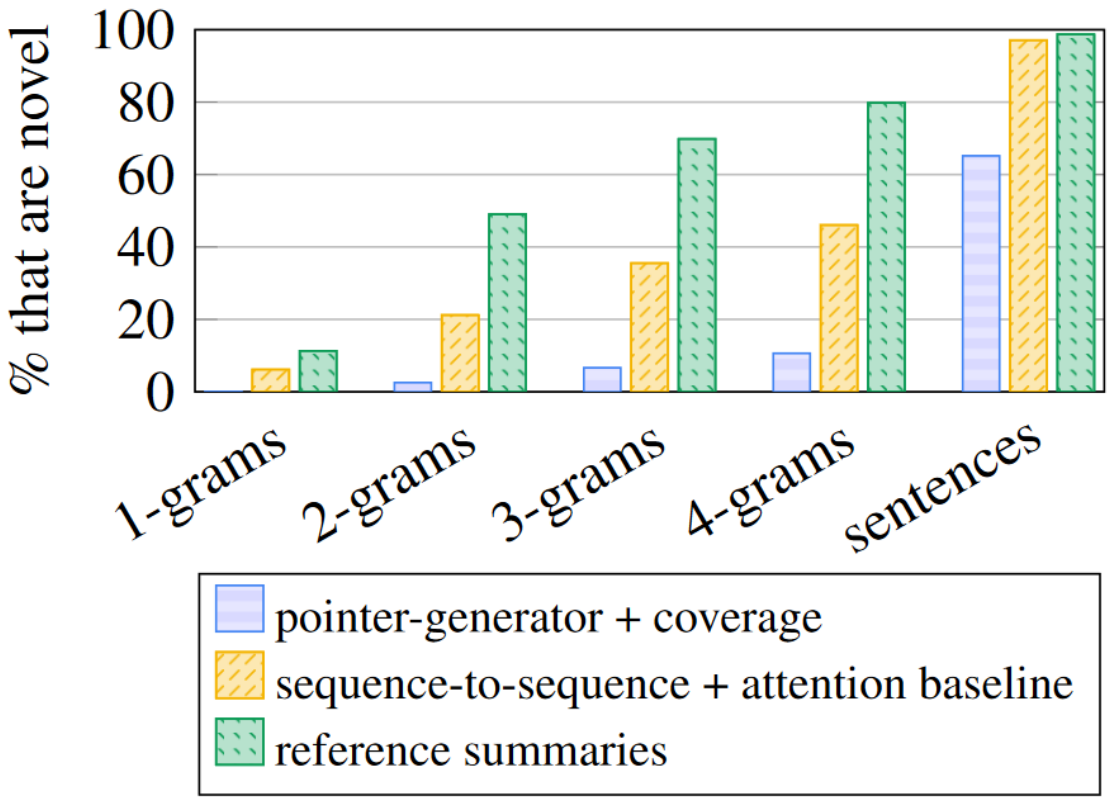
Table 1: ROUGE and METEOR scores on the test set

	ROUGE			METEOR	
	1	2	L	exact match	+ stem/syn/para
abstractive model (Nallapati et al., 2016)*	35.46	13.30	32.65	-	-
seq-to-seq + attn baseline (150k vocab)	30.49	11.17	28.08	11.65	12.86
seq-to-seq + attn baseline (50k vocab)	31.33	11.81	28.83	12.03	13.20
pointer-generator	36.44	15.66	33.42	15.35	16.65
pointer-generator + coverage	**39.53**	**17.28**	**36.38**	17.32	18.72
lead-3 baseline (ours)	40.34	17.70	36.57	20.48	22.21
lead-3 baseline (Nallapati et al., 2017)*	39.2	15.7	35.5	-	-
extractive model (Nallapati et al., 2017)*	39.6	16.2	35.3	-	-

[1] Chin-Yew Lin. 2004b. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out: ACL workshop.
[2] Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In EACL 2014 Workshop on Statistical Machine Translation.

- **Left**: Coverage mechanism eliminates undesirable repetitions
- **Right**: % of novel n-grams generated by proposed model are much smaller



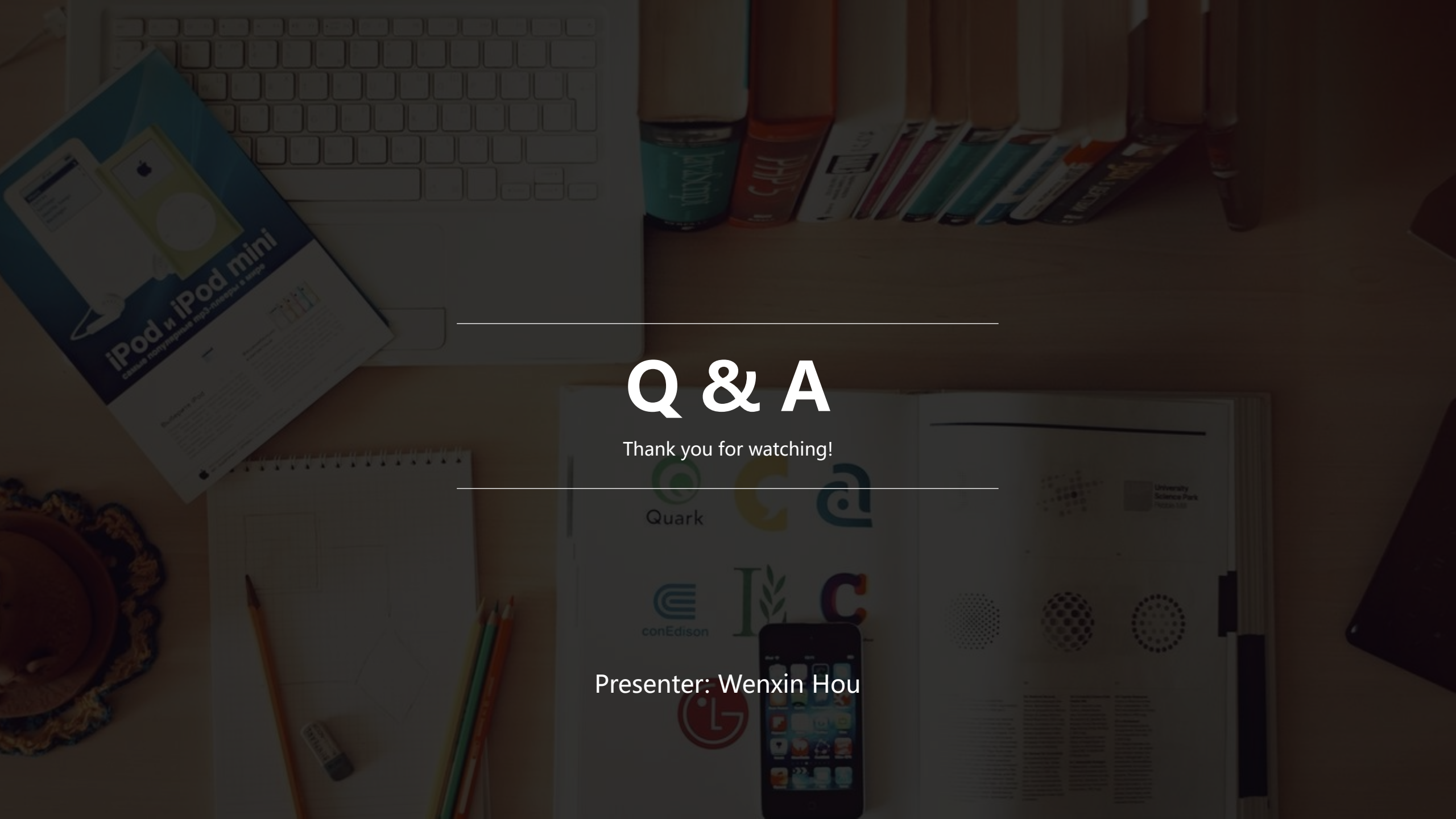


Contribution of the proposed pointer-generator network:

- Reduce inaccuracies and repetition on abstractive summarization
- Outperform the abstractive state-of-the-art result on a challenging long-text dataset

Limitation:

- Attaining higher levels of abstraction remains a question



Q & A

Thank you for watching!

Presenter: Wenxin Hou