Closed-Book Training to Improve Summarization Encoder Memory

Yichen Jiang and Mohit Bansal

www.jiang-yichen.io, www.cs.unc.edu/~mbansal

Abstract

In this paper, we aim to improve the memorization ability of the encoder of a pointer-generator model by adding an additional 'closed-book' decoder without attention/pointer mechanisms.

- Intuition: Such a decoder forces the encoder to be more selective in the information encoded in its memory state because the decoder can't rely on the extra information provided by the attention and possibly copy modules.

We demonstrate our model's superiority to the pointer-generator baseline and prove that our encoder does learn stronger memory representations by showing that our 2-decoder model achieves the following improvements:

- Statistically significant improvements on the ROUGE and METEOR, for both cross-entropy and reinforced setups (and on human evaluation), on CNN/DM and Newsroom datasets.
- Higher scores in a test-only DUC-2002 generalizability setup.
- Extensive analysis shows better results in a memory-ability test, two saliency metrics, and several sanity-check ablations.

Model

Pointer-Generator Baseline: Our abstractive text summarization model is a simple sequence-to-sequence single-layer bidirectional encoder and unidirectional decoder LSTM-RNN, with attention (Bahdanau et al., 2015), pointer-copy, and coverage mechanisms (See et al., 2017). The generation probability is the sum of copy-from-source probability and generate-from-vocabulary probability, weighted by p^t_{gen} .

$$p^t_{gen} = \sigma(U_c c_t + U_s s_t + U_x x_t + b_{ptr})$$

$$P^t_{attn}(w) = p^t_{gen} P^t_{vocab}(w) + (1 - p^t_{gen}) \sum_{i:w_i=w} a^t_i$$

2-Decoder Model: To enhance encoder's memory, we add a closed-book decoder, which is a uni-directional LSTM decoder without attention/pointer layer. The two decoders share a single encoder and word-embedding matrix, while out-of-vocabulary (OOV) words are simply represented as [UNK] for the closed-book decoder. The entire 2-decoder model is represented in Figure 1. During training, we optimize the weighted sum of negative log likelihoods from the two decoders:

$$\mathcal{L}_{XE} = \frac{1}{T}\sum_{t=1}^{T} -((1-\gamma)\log P^t_{attn}(w|x_{1:t}) + \gamma \log P^t_{cbdec}(w|x_{1:t}))$$

where $P^t_{cbdec}(w|x_{1:t})$ is the generation probability from the closed-book decoder.

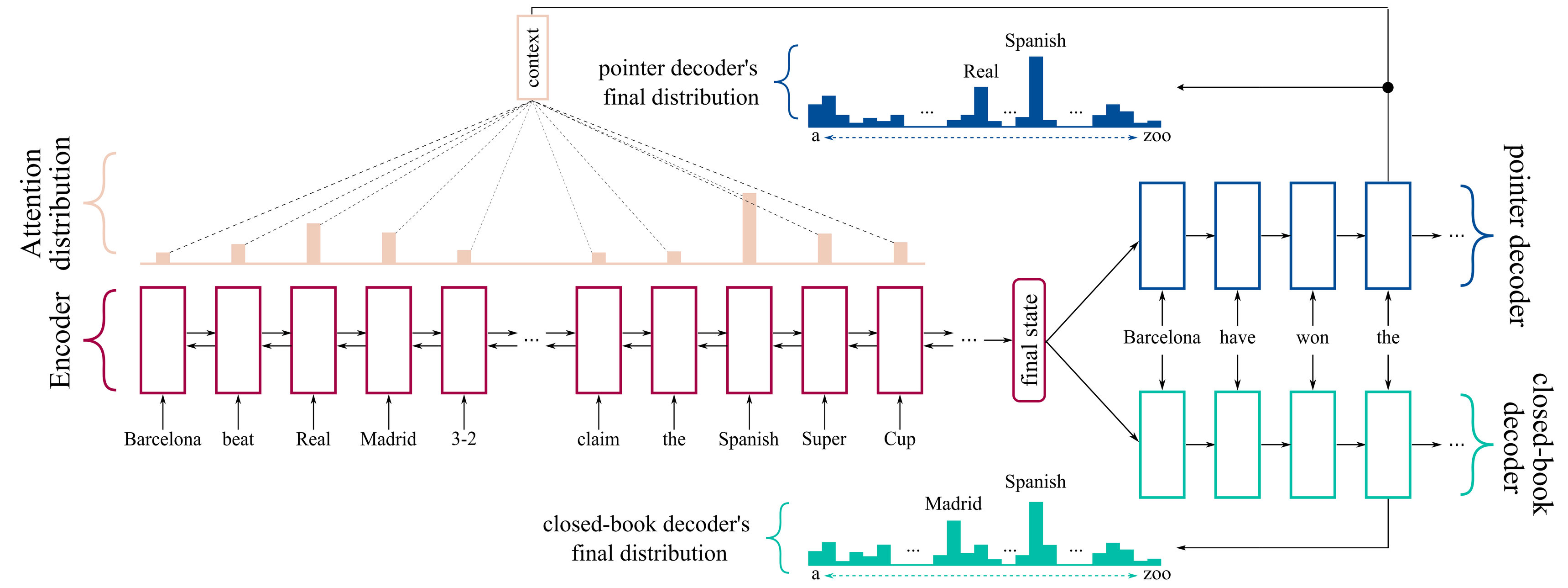

Figure 1: Our 2-decoder model with a pointer decoder and a closed-book decoder sharing a single encoder during training; at inference, we only employ the memory-enhanced encoder and the pointer decoder.

Policy Gradient Reinforce: In order to directly optimize the sentence-level test metrics (as opposed to cross-entropy loss), we use a policy gradient approach where the training objective is to minimize the negative expected reward function. Following Paulus et al. (2018), we also ensure the readability and fluency of the generated summary via a mixed loss function, which is a weighted combination of the cross-entropy and RL losses:

$$L(\theta) = -\mathbb{E}_{w^s \sim p_\theta}[r(w^s)] \qquad \mathcal{L}_{XE+RL} = \lambda \mathcal{L}_{RL} + (1-\lambda)\mathcal{L}_{XE}$$

Results and Ablations

Setup: We use 2 summarization datasets: CNN/Daily Mail and DUC-2002 (test-only transfer setup). Promising initial improvements on Newsroom.

	ROUGE			MTR
	1	2	L	Full
PREVIOUS WORKS				
*(Nallapati16)	35.46	13.30	32.65	
pg (See17)	36.44	15.66	33.42	16.65
OUR MODELS				
pg (baseline)	36.70	15.71	33.74	16.94
pg + cbdec	38.21	16.45	34.70	18.37
RL + pg	37.02	15.79	34.00	17.55
RL + pg + cbdec	**38.58**	**16.57**	**35.03**	**18.86**

Table 1: ROUGE F1 and METEOR scores (*non-coverage*) on **CNN/Daily Mail** test set.

	ROUGE			MTR
	1	2	L	Full
PREVIOUS WORKS				
pg (See17)	39.53	17.28	36.38	18.72
RL* (Paulus17)	39.87	15.82	36.90	
OUR MODELS				
pg (baseline)	39.22	17.02	35.95	18.70
pg + cbdec	40.05	17.66	36.73	19.48
RL + pg	39.59	17.18	36.16	19.70
RL + pg + cbdec	**40.66**	**17.87**	**37.06**	**20.51**

Table 2: ROUGE F1 and METEOR scores (*with-coverage*) on the **CNN/Daily Mail** test set.

	ROUGE			MTR
	1	2	L	Full
pg (See17)	37.22	15.78	33.90	13.69
pg (baseline)	37.15	15.68	33.92	13.65
pg + cbdec	37.59	16.84	34.43	13.82
RL + pg	39.92	16.71	36.13	15.12
RL + pg + cbdec	**41.48**	**18.69**	**37.71**	**15.88**

Table 3: ROUGE F1 and METEOR scores on **DUC-2002** (test-only transfer setup).

	ROUGE		
	1	2	L
$\gamma = 0$	37.73	16.52	34.49
$\gamma = 1/2$	38.09	16.71	34.89
$\gamma = 2/3$	**38.87**	**16.93**	**35.38**
$\gamma = 5/6$	38.21	16.69	34.81
$\gamma = 10/11$	37.99	16.39	34.7

Table 4: Ablation with different 2-decoder mixed-loss ratios, for **CNN/Daily Mail** val set.

Human Evaluation:

Model	Score
2-Decoder Wins	49
Pointer-Generator Wins	31
Non-distinguishable	20

Table 5: Human Evaluation: pairwise comparison between our 2-decoder model and See et al. (2017).

Reference summary:
mitchell moffit and greg brown from asapscience present theories.
different personality traits can vary according to expectations of parents.
beyoncé, hillary clinton and j. k. rowling are all oldest children.

Pointer-Gen baseline:
the kardashians are a strong example of a large celebrity family where the siblings share very different personality traits ......

Pointer-Gen + closed-book decoder:
the kardashians are a strong example of a large celebrity family where the siblings share very different personality traits ......
the personality traits are also supposedly affected by whether parents have high expectations and how strict they were.

Analysis

Memory-Similarity Test: Two forward passes to feed entire article and GT summary to encoder, and compute cosine-similarity between the two final memory states.

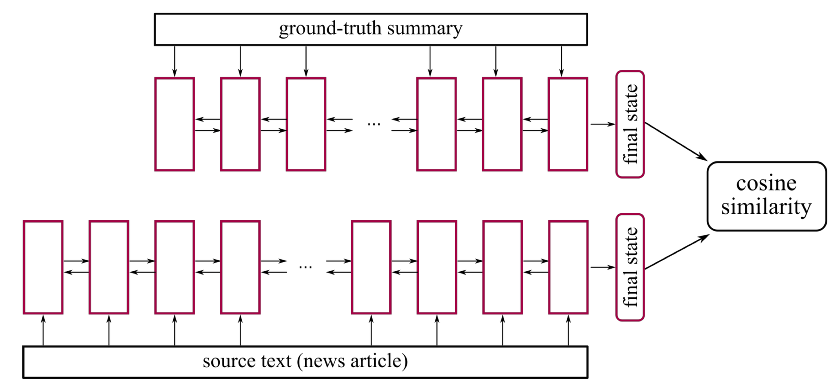

Figure 2: Cosine-similarity between final memory states after reading summary and full doc.

	similarity
pg (baseline)	0.817
pg + cbdec ($\gamma = \frac{1}{2}$)	0.869
pg + cbdec ($\gamma = \frac{2}{3}$)	**0.889**
pg + cbdec ($\gamma = \frac{5}{6}$)	0.872
pg + cbdec ($\gamma = \frac{10}{11}$)	0.860

Table 6: Cosine-similarity between two final memory-states.

Sanity Checks:

	ROUGE		
	1	2	L
FIXED-ENCODER ABLATION			
pg baseline's encoder	37.59	16.27	34.33
2-decoder's encoder	**38.44**	**16.85**	**35.17**
GRADIENT-FLOW-CUT ABLATION			
pg baseline	37.73	16.52	34.49
stop ①	37.72	16.58	34.54
stop ②	**38.35**	**16.79**	**35.13**

Table 7: Fixed-encoder & Gradient-cut ablations.

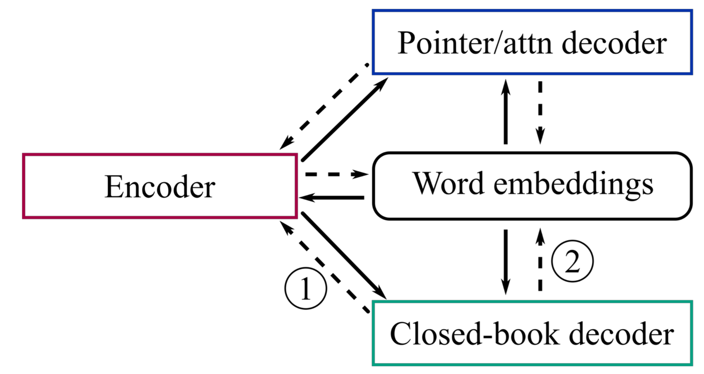

Figure 3: Solid lines = forward pass, dashed lines = gradient flow in backprop; we stop gradient at (1) and (2) for the two ablations.

	ROUGE		
	1	2	L
pg baseline	37.73	16.52	34.49
pg + ptrdec	37.66	16.50	34.47
pg-2layer	37.92	16.48	34.62
pg-big	38.03	16.71	34.84
pg + cbdec	**38.87**	**16.93**	**35.38**

Table 8: Model-capacity sanity-check.

Saliency and Repetition:

	saliency 1	saliency 2
pg (See17)	60.4%	27.95%
our pg baseline	59.6%	28.95%
pg + cbdec	62.1%	29.97%
RL + pg	62.5%	30.17%
RL + pg + cbdec	**66.2%**	**31.40%**

Table 9: Saliency scores based on cloze blank-filling task & keyword-detection (Pasunuru & Bansal, 2018).

	3-gram	4-gram	5-gram	sent
pg (baseline)	13.20%	12.32%	11.60%	8.39%
pg + cbdec	**9.66%**	**9.02%**	**8.55%**	**6.72%**

Table 10: Percentage of repeated 3,4,5-grams & sentences in generated summaries.

Original Text (truncated): a family have claimed the body of an infant who was discovered deceased and buried on a sydney beach last year , in order to give her a proper funeral . on november 30 , 2014 , two young boys were playing on maroubra beach when they uncovered the body of a baby girl buried under 30 centimetres of sand . now locals filomena d'alessandro and bill green have claimed the infant 's body in order to provide her with a fitting farewell . 'we're local and my husband is a police officer and he's worked with many of the officers investigating it ,' ms d'alessandro told daily mail australia . scroll down for video . a sydney family have claimed the body of a baby girl who was found buried on maroubra beach (pictured) on november 30 , 2014 . filomena d'alessandro and bill green have claimed the infant 's remains , who they have named lily grace , in order to provide her with a fitting farewell . ' above all as a mother i wanted to do something for that little girl , ' she added . since january the couple , who were married last year and have three children between them , have been trying to claim the baby after they heard police were going to give her a ' destitute burial ' ...

Reference summary:
sydney family claimed the remains of a baby found on maroubra beach .
filomena d'alessandro and bill green have vowed to give her a funeral .
the baby 's body was found by two boys , buried in sand on november 30 . the infant was found about 20-30 metres from the water 's edge .
police were unable to identify the baby girl or her parents .

Pointer-Generator baseline:
a sydney family have *claimed the body of a baby girl* was found buried on maroubra beach on november 30 , 2014 .
locals filomena d'alessandro and bill green have *claimed the infant 's body* in order to provide her with a fitting farewell .
now locals have *claimed the infant 's body* in order to provide her with a fitting farewell .

Pointer-Generator + closed-book decoder:
two young boys were playing on maroubra beach when they uncovered the body of a baby girl buried under 30 centimetres of sand .
now locals filomena d'alessandro and bill green have claimed the infant 's body in order to provide her with a fitting farewell .
`above all as a mother i wanted to do something for that little girl ,' she added .

Acknowledgments:

We thank the reviewers for their helpful comments. This work was supported by DARPA (YFA17-D17AP00022), Google Faculty Research Award, Bloomberg Data Science Research Grant, Nvidia GPU awards, and Amazon AWS. The views contained in this article are those of the authors and not of the funding agency.

References:

- Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *ACL*.
- Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *ICLR*.
- Ramakanth Pasunuru and Mohit Bansal. 2018. Multi-reward reinforced summarization with saliency and entailment. In NAACL.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.