Robust Neural Abstractive Summarization Systems and Evaluation against Adversarial Information

Lisa Fan[1], Dong Yu[2], Lu Wang[1]

[1] Northeastern University {lisafan, luwang}@ccs.neu.edu

[2] Tencent AI Lab dyu@tencent.com

Introduction

Sequence-to-sequence (seq2seq) neural models[1] have been actively investigated for abstractive summarization. Nevertheless, existing neural abstractive systems frequently generate factually incorrect summaries and are vulnerable to adversarial information, suggesting a crucial lack of semantic understanding.

We propose a novel semantic-aware neural abstractive summarization model that learns to generate high quality summaries through semantic interpretation over salient content. A novel evaluation scheme with adversarial samples is introduced to measure how well a model identifies off-topic information, where our model yields significantly better performance than the popular pointer-generator summarizer[2]. Human evaluation also confirms that our system summaries are uniformly more informative and faithful as well as less redundant than the seq2seq model.

Models

Shared Decoder

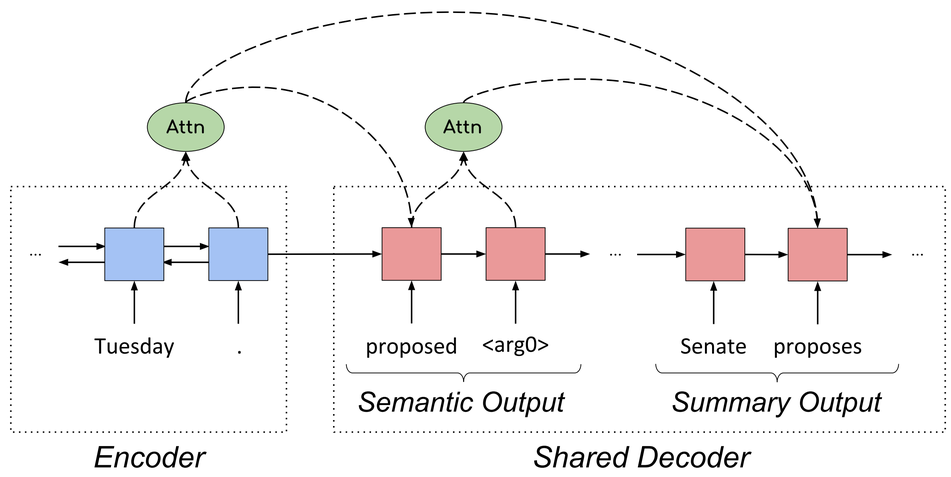


Figure 1. The shared decoder model uses a single LSTM encoder and a single LSTM decoder to generate the semantic roles followed by the summary. During semantic role generation, the decoder attends over the input. During summary generation, the decoder attends over both the input and the generated semantic roles.

Separate Decoder

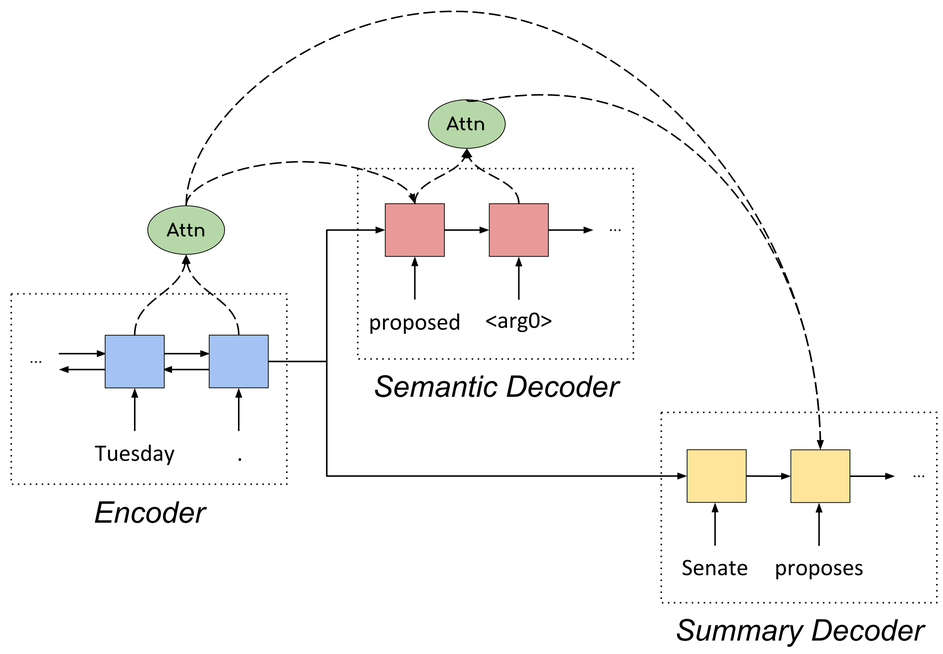


Figure 2. The separate decoder model uses a single LSTM encoder and two LSTM decoders to generate either the semantic roles or the summary. The semantic decoder attends over the input. The summary decoder attends over both the input and the generated semantic roles.

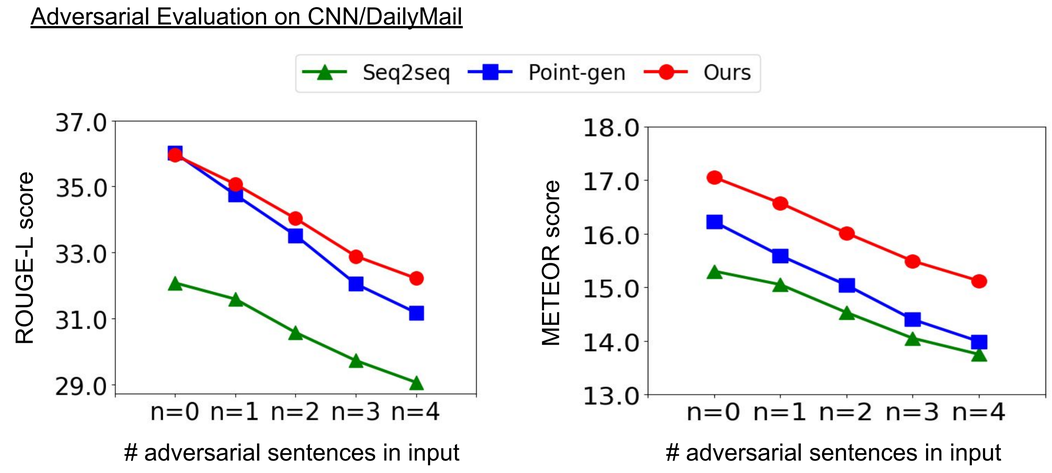
References

[1] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

[2] Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1.

[3] Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 708-719.

Adversarial Evaluation



Example Adversarial Input: For years Joe DiMaggio was always introduced at Yankee Stadium as "baseball's greatest living player." But with his memory joining those of Babe Ruth, Lou Gehrig, Mickey Mantle and Miller Huggins. Canada's Minister of Defense resigned today, a day after an army official testified that top military officials had altered documents to cover up responsibility for the beating death of a Somali teen-ager at the hands of Canadian peacekeeping troops in 1993. Defense minister David Collenette insisted that his resignation had nothing to do with the Somalia scandal. Ted Williams was the first name to come to mind, and he's the greatest living hitter. Collenette told reporters in Ottawa that he was stepping down because he had written a letter to help a constituent win an immigration case.

Seq2seq: George Vecsey sports of The Times column on New York State's naming of late baseball legend Joe DiMaggio as "baseball's greatest living player," but with his memory joining those of Babe Ruth, Lou Gehrig, Mickey Mantle and Miller dens.

Our Model: Former Canadian Defense Min David Collenette resigns day after army official testifies that top military officials altered documents to cover up responsibility for beating death of Somali teen-ager at hands of Canadian peacekeeping troops in 1993.

Figure 3. Example adversarial input and outputs. We add a small number of off-topic sentences into a test article. The intuition is that if a summarization system truly understands the salient entities and events, it would ignore unrelated sentences. The adversarial sentences are highlighted in red.

Automatic Evaluation

	CNN/DailyMail						
	ROUGE			METEOR		Dens.	Red.
	R-1	R-2	R-L	exact	full		
HUMAN	-	-	-	-	-	3.40	0.06
LEAD-2	38.35	15.74	31.48	14.44	15.62	59.19	0.33
TEXTRANK	33.78	12.41	28.75	12.73	13.94	36.44	7.94
Abstractive Comparisons							
SEQ2SEQ	32.14	12.33	29.41	13.66	14.65	8.46	11.92
POINT-GEN	35.60	15.24	32.43	15.50	16.54	16.80	10.77
Our Semantic-Aware Models							
DEC$_{share}$	35.44	14.18	32.28	14.80	15.84	12.42	5.97
+ MHA	**36.31**	14.96	**33.21**	**15.56**	**16.62**	13.03	**5.20**
DEC$_{sep}$	35.31	13.97	32.23	14.81	15.86	**11.55**	5.21
+ MHA	36.07	**15.09**	33.08	15.25	16.29	12.93	5.65

Table 1. Results on the CNN/Daily Mail dataset. For our model, we display four variants, based on shared or separate decoder, and with or without multi-head attention (MHA). In addition to ROUGE and METEOR, we also display extractive density (Dens.) and redundancy (Red.) (lower scores are preferred). Best performing amongst our models are in bold.

$$\text{DENSITY}(A,S) = \frac{1}{|S|}\sum_{f \in F(A,S)} |f|^2$$

$$\text{REDUNDANCY}(S) = \frac{1}{|S|}\sum_{f' \in F'(S)} (\#f' \times |f'|)^2$$

Density measures how much the summary reuses the input article verbatim.[3]

Redundancy measures how much the summary repeats itself.

Human Evaluation

	NON-RED.	FLUENCY	FAITH.	INFORM.
HUMAN	1.4 ± 0.7	1.5 ± 0.7	1.5 ± 0.8	1.6 ± 0.8
SEQ2SEQ	2.0 ± 0.8	2.4 ± 0.7	2.1 ± 0.8	2.4 ± 0.7
OURS	1.6 ± 0.7	2.1 ± 0.8	1.9 ± 0.7	2.0 ± 0.7

Table 2. We compare the outputs of our shared decoder model, the traditional seq2seq model, and the human reference summaries. We asked human judges to read the article and rank the summaries against each other based on non-redundancy, fluency, faithfulness to input, and informativeness (whether the summary delivers the main points of the article). The mean (± std. dev.) for the rankings is shown (lower is better).