Improving Abstractive Sentence Summarization from the Encoder and Decoder Sides

Qingyu Zhou

2018,03,15

Outline

Background

From the Encoder Side

From the Decoder Side

Reference

Background

- Task Definition
- Related Work

Automatic Text Summarization

Automatic summarization is the process of shortening a text document with software, in order to create a summary with the major points of the original document. (from Wikipedia)

By the input granularity:

- sentence summarization
- single-document summarization
- multi-document summarization

By the output production method:

- extractive summarization
- abstractive summarization

From other perspectives:

- email, stream summarization
- query / topic focused summarization
- etc.

Abstract Sentence Summarization

Similar to Document Summarization:

- Extractive
- **Abstractive**

Sentence Summarization:

- Input: A long sentence
- Output: Its shorter version
- Also known as *Sentence Compression*
- The biggest dataset is essentially for *Headline Generation*

Related Work

The very basic models:

- Sequence-to-Sequence framework [Sutskever et al., 2014]
- Attention Mechanism [Bahdanau et al., 2015]

On Abstractive Sentence Summarization task:

- A Neural Attention Model for Abstractive Sentence Summarization [Rush et al., 2015]
- Abstractive Sentence Summarization with Attentive Recurrent Neural Networks [Chopra et al., 2016]
- Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond [Nallapati et al., 2016]

S2S with Attention

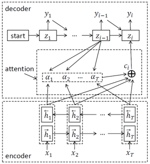
- Encoder : BiGRU [Cho et al., 2014]

$$z_i = \sigma(\mathbf{W}_z[x_i, h_{i-1}]) \tag{1}$$
$$r_i = \sigma(\mathbf{W}_r[x_i, h_{i-1}]) \tag{2}$$
$$\widetilde{h}_i = \tanh(\mathbf{W}_h[x_i, r_i \odot h_{i-1}]) \tag{3}$$
$$h_i = (1 - z_i) \odot h_{i-1} + z_i \odot \widetilde{h}_i \tag{4}$$

- Decoder:

$$p(y_j | y_{<j}, H) = g(s_j, y_{j-1}, c_j) \tag{5}$$
$$s_t = \mathrm{GRU}(w_{t-1}, c_{t-1}, s_{t-1}) \tag{6}$$

- Attention Mechanism:



$$e_{t,i} = v_a^\top \tanh(\mathbf{W}_a s_{t-1} + \mathbf{U}_a h_i) \tag{7}$$
$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{i=1}^{n} \exp(e_{t,i})} \tag{8}$$
$$c_t = \sum_{i=1}^{n} \alpha_{t,i} h_i \tag{9}$$

Rush et al. [2015]

Rush, Alexander M., et al. "A Neural Attention Model for Sentence Summarization." EMNLP. 2015.

- NNLM

$$
\begin{aligned}
p(\mathbf{y}_{i+1}|\mathbf{y}_c, \mathbf{x}; \theta) &\propto \exp(\mathbf{V}\mathbf{h} + \mathbf{W}\mathrm{enc}(\mathbf{x}, \mathbf{y}_c)), \\
\tilde{\mathbf{y}}_c &= [\mathbf{E}\mathbf{y}_{i-C+1}, \ldots, \mathbf{E}\mathbf{y}_i], \\
\mathbf{h} &= \tanh(\mathbf{U}\tilde{\mathbf{y}}_c).
\end{aligned}
$$

- Attention-Based Encoder

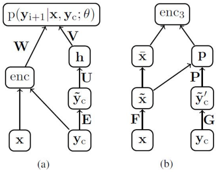


Figure 3: (a) A network diagram for the NNLM decoder with additional encoder element. (b) A network diagram for the attention-based encoder enc_3.

$$
\begin{aligned}
\mathrm{enc}_3(\mathbf{x}, \mathbf{y}_c) &= \mathbf{p}^\top \bar{\mathbf{x}}, \\
\mathbf{p} &\propto \exp(\tilde{\mathbf{x}}\mathbf{P}\tilde{\mathbf{y}}'_c), \\
\tilde{\mathbf{x}} &= [\mathbf{F}\mathbf{x}_1, \ldots, \mathbf{F}\mathbf{x}_M], \\
\tilde{\mathbf{y}}'_c &= [\mathbf{G}\mathbf{y}_{i-C+1}, \ldots, \mathbf{G}\mathbf{y}_i], \\
\forall i \quad \bar{\mathbf{x}}_i &= \sum_{q=i-Q}^{i+Q} \tilde{\mathbf{x}}_i / Q.
\end{aligned}
$$

Chopra et al. [2016]

Chopra, Sumit, Michael Auli, and Alexander M. Rush. "Abstractive sentence summarization with attentive recurrent neural networks." NAACL. 2016.

- Based on Rush et al. [2015]
- CNN encoder + RNN decoder

Nallapati et al. [2016]

Nallapati, Ramesh, et al. "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond." CoNLL. 2016.

- Full RNN based s2s with attention
- Feature-rich encoder
- Hierarchical encoder for document level summarization

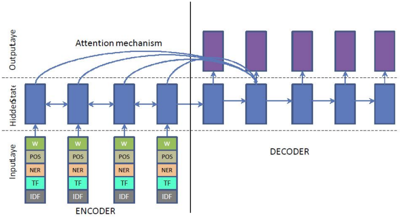


Figure 1: Feature-rich-encoder: We use one embedding vector each for POS, NER tags and discretized TF and IDF values, which are concatenated together with word-based embeddings as input to the encoder.

Evaluation

Human Evaluation

Automatic Evaluation

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [Lin, 2004]
  - ROUGE-N: n-gram overlap between system output and reference
    - ROUGE-1: unigram overlap
    - ROUGE-2: bigram overlap
  - ROUGE-L: LCS (longest common subsequence)

From the Encoder Side

Selective Encoding for Abstractive Sentence Summarization
Qingyu Zhou, Nan Yang, Furu Wei and Ming Zhou
Proc. ACL 2017.

SEASS – Motivation

Attention-based Sequence-to-Sequence (s2s) framework

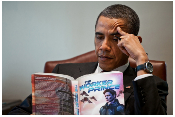
- RNN based encoder and decoder
- Encoder: encode input tokens to a list of vectors
- Decoder: decode the encoded information to produce output tokens
- Attention: Soft-alignment between input and output words

However, in abstractive sentence summarization, there is no explicit alignment relationship except for the extracted words.

SEASS – Motivation

How does human do summarization?

- Read sentence
- Select important contents
- Write summary



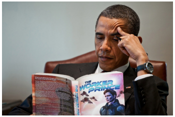
the **sri lankan** government on wednesday announced the **closure** of **government schools** with **immediate effect** as a **military campaign** against **tamil separatists escalated** in the north of the country .

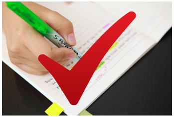
sri lanka closes schools as war escalates

SEASS – Motivation

How does human do summarization?

- Read sentence
- Select important contents
- Write summary





the **sri lankan** government on wednesday announced the **closure** of **government schools** with **immediate effect** as a **military campaign** against **tamil separatists escalated** in the north of the country .

sri lanka closes schools as war escalates

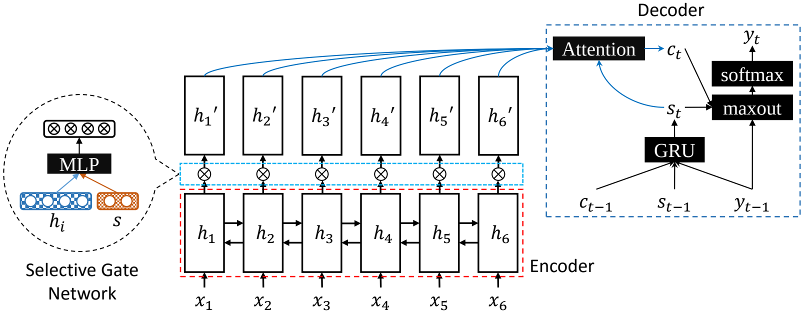
SEASS

- Explicitly modeling the importance of each word
- Build a tailored representation with the proposed selective mechanism for the abstractive sentence summarization task.
- **S**elective **E**ncoding for **A**bstractive **S**entence **S**ummarization, **SEASS**

the **sri lankan** government on wednesday announced the **closure** of **government schools** with **immediate effect** as a **military campaign** against **tamil separatists escalated** in the north of the country .

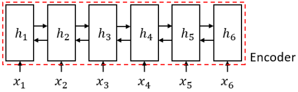
sri lanka closes schools as war escalates

SEASS – Overview



SEASS – Encoder and Decoder

- Encoder: BiGRU



$$\vec{h}_i = \mathrm{GRU}(x_i, \vec{h}_{i-1}) \tag{10}$$

$$\overleftarrow{h}_i = \mathrm{GRU}(x_i, \overleftarrow{h}_{i+1}) \tag{11}$$

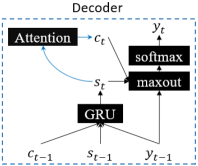
$$h_i = [\vec{h}_i ; \overleftarrow{h}_i] \tag{12}$$

- Decoder

$$s_t = \mathrm{GRU}(c_{t1}, s_{t1}, y_{t1}) \tag{13}$$

$$p(y_t | y_{<t}, X) = g(y_{t1}, s_t, c_t) \tag{14}$$

- Attention Mechanism



$$e_{t,i} = v_a^\top \tanh(\mathbf{W}_a s_{t-1} + \mathbf{U}_a h'_i) \tag{15}$$

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{i=1}^{n} \exp(e_{t,i})} \tag{16}$$

$$c_t = \sum_{i=1}^{n} \alpha_{t,i} h'_i \tag{17}$$

SEASS – Selective Encoding

- Sentence Meaning Representation

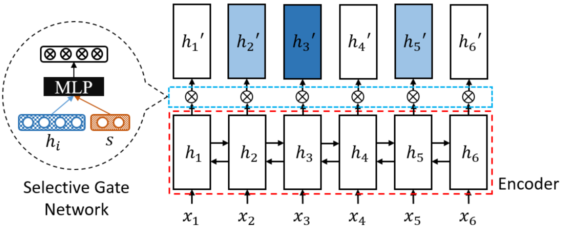
$$s = \begin{bmatrix} \overleftarrow{h}_1 \\ \overrightarrow{h}_n \end{bmatrix} \tag{18}$$

- Word Importance

$$sGate_i = \sigma(\mathbf{W}_s h_i + \mathbf{U}_s s + b) \tag{19}$$

- Tailored Representation

$$h_i' = h_i \odot sGate_i \tag{20}$$



Datasets

- Giga: English Gigaword [Napoles et al., 2012]
- DUC: DUC 2004 summarization task data set [Over and Yen, 2004]
- MSR-ATC: Microsoft Research Abstractive Text Compression [Toutanova et al., 2016]

Data Set	Giga	DUC†	MSR†
#(sent)	3.99M	500	785
#(sentWord)	125M	17.8K	29K
#(summWord)	33M	20.9K	85.9K
#(ref)	1	4	3-5
AvgInputLen	31.35	35.56	36.97
AvgSummLen	8.23	10.43	25.5

Table: Data statistics for the English Gigaword, DUC 2004 and MSR-ATC datasets. $\#(x)$ denotes the number of x, e.g., #(ref) is the number of reference summaries of an input sentence. AvgInputLen is the average input sentence length and AvgSummLen is the average summary length. †DUC 2004 and MSR-ATC datasets are for test purpose only.

Baseline

- ABS Rush et al. [2015] use an attentive CNN encoder and NNLM decoder to do the sentence summarization task. We trained this baseline model with the released code and evaluate it with our internal English Gigaword test set and MSR-ATC test set.
- ABS+ Based on ABS model, Rush et al. [2015] further tune their model using DUC 2003 dataset, which leads to improvements on DUC 2004 test set.
- CAs2s As an extension of the ABS model, Chopra et al. [2016] use a convolutional attention-based encoder and RNN decoder, which outperforms the ABS model.
- Feats2s Nallapati et al. [2016] use a full RNN sequence-to-sequence encoder-decoder model and add some features to enhance the encoder, such as POS tag, NER, and so on.
- Luong-NMT Neural machine translation model of Luong et al. [2015] with two-layer LSTMs for the encoder-decoder with 500 hidden units in each layer implemented in Chopra et al. [2016].
- s2s+att We also implement a sequence-to-sequence model with attention as our baseline and denote it as "s2s+att".

SEASS – English Gigaword

Models	RG-1	RG-2	RG-L
ABS (beam)‡	29.55⁻	11.32⁻	26.42⁻
ABS+ (beam)‡	29.76⁻	11.88⁻	26.96⁻
Feats2s (beam)‡	32.67⁻	15.59⁻	30.64⁻
CAs2s (greedy)‡	33.10⁻	14.45⁻	30.25⁻
CAs2s (beam)‡	33.78⁻	15.97⁻	31.15⁻
Luong-NMT (beam)‡	33.10⁻	14.45⁻	30.71⁻
s2s+att (greedy)	33.18⁻	14.79⁻	30.80⁻
s2s+att (beam)	34.04⁻	15.95⁻	31.68⁻
SEASS (greedy)	35.48	16.50	32.93
SEASS (beam)	**36.15**	**17.54**	**33.63**

Table: Full length ROUGE F1 evaluation results on the English Gigaword test set used by Rush et al. [2015]. RG in the Table denotes ROUGE. Results with ‡ mark are taken from the corresponding papers. The superscript ⁻ indicates that our SEASS model with beam search performs significantly better than it as given by the 95% confidence interval in the official ROUGE script.

SEASS – DUC 2004

Models	RG-1	RG-2	RG-L
ABS (beam)‡	26.55^{-}	7.06^{-}	22.05^{-}
ABS+ (beam)‡	28.18^{-}	8.49^{-}	23.81^{-}
Feats2s (beam)‡	28.35^{-}	9.46	24.59^{-}
CAs2s (greedy)‡	29.13	7.62^{-}	23.92^{-}
CAs2s (beam)‡	28.97	8.26^{-}	24.06^{-}
Luong-NMT (beam)‡	28.55	8.79^{-}	24.43^{-}
s2s+att (greedy)	27.03^{-}	7.89^{-}	23.80^{-}
s2s+att (beam)	28.13	9.25	24.76
SEASS (greedy)	28.68	8.55	25.04
SEASS (beam)	**29.21**	**9.56**	**25.51**

Table: ROUGE recall evaluation results on DUC 2004 test set. All these models are tested using beam search. Results with ‡ mark are taken from the corresponding papers. The superscript $^{-}$ indicates that our SEASS model performs significantly better than it as given by the 95% confidence interval in the official ROUGE script.

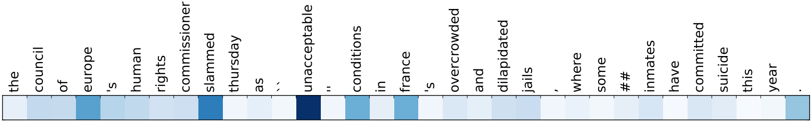
SEASS – MSR-ATC

Models	RG-1	RG-2	RG-L
ABS (beam)	20.27^{-}	5.26^{-}	17.10^{-}
s2s+att (greedy)	15.15^{-}	4.48^{-}	13.62^{-}
s2s+att (beam)	22.65^{-}	9.61^{-}	21.39^{-}
SEASS (greedy)	19.77	6.44	17.36
SEASS (beam)	**25.75**	**10.63**	**22.90**

Table: Full length ROUGE F1 evaluation on our internal English Gigaword test data. The superscript $^{-}$ indicates that our SEASS model performs significantly better than it as given by the 95% confidence interval in the official ROUGE script.

Saliency Heat Map of Selective Gate

- First-Derivative Saliency method in *Visualizing and Understanding Neural Models in NLP* [Li et al., 2016] .
- To visualize the high dimensional first derivative, we instead draw its Euclidean norm .



First derivative heat map of the output with respect to the selective gate. The important words are selected in the input sentence, such as "europe", "slammed" and "unacceptable". The output summary of our system is "*council of europe slams french prison conditions*" and the true summary is "*council of europe again slams french prison conditions*".

Conclusion

- We propose Selective Encoding for abstractive sentence summarization to model the important of words in given sentence.
- The experimental results on three data sets demonstrates the efficacy of SEASS.
- For future work, we will apply this selective mechanism to other tasks such sentiment analysis.

From the Decoder Side

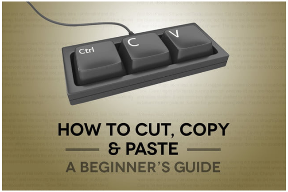
Sequential Copying Networks
Qingyu Zhou, Nan Yang, Furu Wei and Ming Zhou
Proc. AAAI 2018[1].

[1] Conference proceeding not available yet. Check out our poster at https://res.qyzhou.me/AAAI2018_poster.pdf

Sequential Copying Networks: Motivation

How does human do summarization?

- Copy

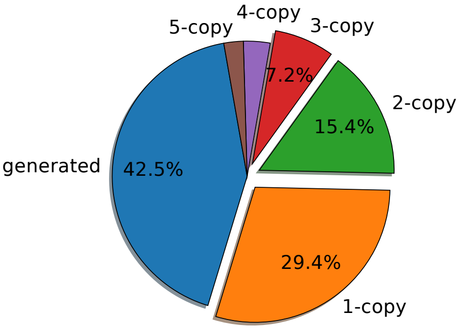


the chilean foreign ministry announced wednesday that it will revise the security regime at its embassies overseas.

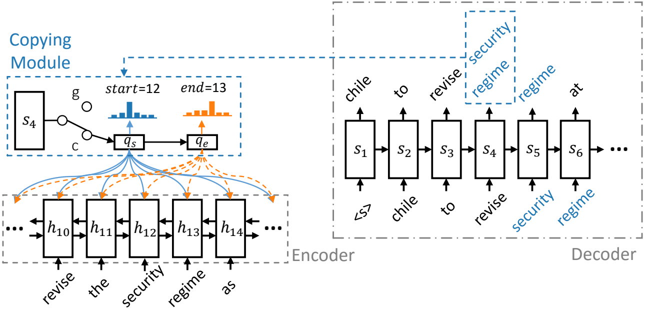
chile to revise security regime at embassies overseas

Sequential Copying Networks: Motivation

Percentage of generated and copied words in the training set of abstractive sentence summarization



Sequential Copying Networks (SeqCopyNet)



SeqCopyNet – Span Prediction

- copy switch gate network
  - produce a probability of copying
- pointer network
  - find the span that should be copied
- copy state transducer
  - assist the pointer network

SeqCopyNet – Span Prediction

- decoder memory vector m_t:

$$m_t = \begin{bmatrix} y_{t-1} \\ s_t \\ c_t \end{bmatrix} \tag{21}$$

- probability of copying

$$p_c = \mathcal{G}(m_t) \tag{22}$$
$$p_g = 1 - p_c \tag{23}$$

- start query vector q_s:

$$q_s = \tanh(\mathbf{W}_s m_t + b) \tag{24}$$

SeqCopyNet – Span Prediction

- span start position copy_s :

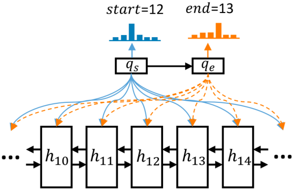
$$e_{s,i} = v_p^\top \tanh(\mathbf{W}_p q_s + \mathbf{U}_p h_i) \tag{25}$$

$$\alpha_{s,i} = \frac{\exp(e_{s,i})}{\sum_{i=1}^{n} \exp(e_{s,i})} \tag{26}$$

$$\text{copy}_s = \arg\max_i \alpha_{s,i} \tag{27}$$

$$p_{\text{copy\_s}} = \alpha_{s,\text{copy}_s} \tag{28}$$

$$c_s = \sum_{i=1}^{n} \alpha_{s,i} h_i \tag{29}$$



SeqCopyNet – Span Prediction

- end query vector q_e:

$$cst = \tanh(\mathbf{W}_e m_t + b) \tag{30}$$

$$q_e = \mathrm{GRU}(cst, c_s) \tag{31}$$

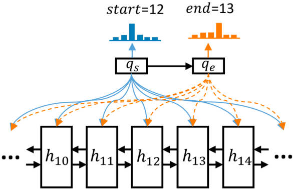
- span end position $\mathrm{copy_e}$:

$$e_{e,i} = v_p^\top \tanh(\mathbf{W}_p q_e + \mathbf{U}_p h_i) \tag{32}$$

$$\alpha_{e,i} = \frac{\exp(e_{e,i})}{\sum_{i=1}^{n} \exp(e_{e,i})} \tag{33}$$

$$\mathrm{copy_e} = \arg\max_i \alpha_{e,i} \tag{34}$$

$$p_{\mathrm{copy\_e}} = \alpha_{e,\mathrm{copy_e}} \tag{35}$$



SeqCopyNet – Gigaword

Models	Test set of Zhou et al. [2017] RG-1	RG-2	RG-L	Internal test set[2] RG-1	RG-2	RG-L
ABS‡	37.41-	15.87-	34.70-	-	-	-
s2s+att (greedy)	46.21	24.02	43.30	45.46	22.83	42.66
s2s+att (beam)	47.08	25.11	43.81	46.54	24.18	43.55
NMT + UNK_PS (greedy)	45.64	23.38	42.67	45.21	23.01	42.38
NMT + UNK_PS (beam)	47.05	24.82	43.87	46.52	24.41	43.58
SEASS (greedy)‡	45.27	22.88	42.20	-	-	-
SEASS (beam)‡	46.86	24.58	43.53	-	-	-
SeqCopyNet (greedy)	46.51	24.14	43.20	46.08	23.99	43.26
SeqCopyNet (beam)	**47.27**	25.07	**44.00**	**47.13**	**24.93**	**44.06**

[2] We released this test set.

SeqCopyNet – Case Study on Gigaword

Input:	david ortiz homered and scored three times , including the go-ahead run in the eighth inning , as the boston *red sox* beat the toronto *blue jays 10-9* in the american league on tuesday .
Reference:	david ortiz helps red sox beat blue jays 10-9
SingleCopy:	ortiz homers as red sox beat blue jays
SeqCopyNet:	[*red sox*] beat [*blue jays 10-9*]
Input:	*guyana 's president cheddi jagan* , a long-time marxist turned free - marketeer , died here early thursday , an embassy spokeswoman said . he was 78 .
Reference:	guyana 's president cheddi jagan marxist turned marketeer dies at 78
SingleCopy:	guyana 's president jagan dies at 78
SeqCopyNet:	[*guyana 's president cheddi jagan*] dies at 78
Input:	china topped myanmar 's marine *product exporting countries annually* in the past decade among over 40 's , the local voice weekly quoted the marine products producers and exporters association as reporting sunday .
Reference:	china tops myanmars marine product exporting countries in past
SingleCopy:	china tops myanmar 's marine product export
SeqCopyNet:	china tops myanmar 's marine [*product exporting countries annually*]

SeqCopyNet – Question Generation

Model	Dev set	Test set
PCFG-Trans‡	9.28	9.31
s2s+att‡	3.01	3.06
NQG‡	10.06	10.13
NQG+‡ (single copy)	12.30	12.18
SeqCopyNet	**13.13**	**13.02**

SeqCopyNet – Case Study on QG

I:	*peyton manning* became the first quarterback ever to lead two different teams to multiple super bowls .
G:	how many teams has manning played for that reached the super bowl , while he was on their team ?
O:	how many teams did [*peyton manning*] lead ?
I:	it is conjectured that a progressive *decline in hormone levels* with age is partially responsible for weakened immune responses in aging individuals .
G:	what is partially responsible for weakened immune response in older individuals ?
O:	what is [*responsible for weakened immune*] responses in aging individuals ?
I:	*the sarah jane adventures*, starring elisabeth sladen who reprised her role as investigative journalist sarah jane smith, was developed by cbbc; a special aired on new year's day 2007 and a full series began on 24 september 2007.
G:	when did the sarah jane series begin ?
O:	on what date did [*the sarah jane adventures*] begin ?

Conclusion

- SeqCopyNet enables multi-word span copying, and can be integrated with seq2seq framework
- SeqCopyNet is good at detecting boundaries, such as named entity
- We release a new abstractive sentence summarization test set
- Future work: apply SeqCopyNet to other tasks such as dialogue generation

Thank you

Q&A

qyzhou@hit.edu.cn

Reference

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of 3rd International Conference for Learning Representations*, San Diego, 2015.

Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

Sumit Chopra, Michael Auli, and Alexander M. Rush. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, San Diego, California, June 2016. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Ça glar Gulçehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, 2016.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain, 2004.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, AKBC-WEKEX '12, pages 95–100, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

Paul Over and James Yen. Introduction to duc-2001: an intrinsic evaluation of generic news text summarization systems. In *Proceedings of DUC 2004 Document Understanding Workshop, Boston*, 2004.

Kristina Toutanova, Chris Brockett, Ke M. Tran, and Saleema Amershi. A dataset and evaluation metrics for abstractive compression of sentences and short paragraphs. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 340–350, Austin, Texas, November 2016. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. Visualizing and understanding neural models in nlp. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California, June 2016. Association for Computational Linguistics.

Qingyu Zhou, Nan Yang, Furu Wei, and Ming Zhou. Selective encoding for abstractive sentence summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1095–1104, Vancouver, Canada, July 2017. Association for Computational Linguistics. URL `http://aclweb.org/anthology/P17-1101`.