Countering the Effects of Lead Bias in News Summarization via Multi-Stage Training and Auxiliary Losses

Matt Grenander[*], **Yue Dong**[*]
McGill University / MILA
{matthew.grenander, yue.dong2}@mail.mcgill.ca

Jackie C. K. Cheung
McGill University / MILA
jcheung@cs.mcgill.ca

Annie Louis
University of Edinburgh
Alan Turing Institute
alouis@inf.ed.ac.uk

Abstract

Sentence position is a strong feature for news summarization, since the lead often (but not always) summarizes the key points of the article. In this paper, we show that recent neural systems excessively exploit this trend, which although powerful for many inputs, is also detrimental when summarizing documents where important content should be extracted from later parts of the article. We propose two techniques to make systems sensitive to the importance of content in different parts of the article. The first technique employs 'unbiased' data; i.e., randomly shuffled sentences of the source document, to pretrain the model. The second technique uses an auxiliary ROUGE-based loss that encourages the model to distribute importance scores throughout a document by mimicking sentence-level ROUGE scores on the training data. We show that these techniques significantly improve the performance of a competitive reinforcement learning based extractive system, with the auxiliary loss being more powerful than pretraining.

1 Introduction

Extractive summarization remains a simple and fast approach to produce summaries which are grammatical and accurately represent the source text. In the news domain, these systems are able to use a dominant signal: the position of a sentence in the source document. Due to journalistic conventions which place important information early in the articles, the lead sentences often contain key information. In this paper, we explore how systems can look beyond this simple trend.

Naturally, automatic systems have all along exploited position cues in news as key indicators of important content (Schiffman et al., 2002; Hong and Nenkova, 2014; Liu, 2019). The 'lead' baseline is rather strong in single-document news summarization (Brandow et al., 1995; Nenkova, 2005), with automatic systems only modestly improving the results. Nevertheless, more than 20-30% of summary-worthy sentences come from the second half of news documents (Nallapati et al., 2016; Kedzie et al., 2018), and the lead baseline, as shown in Table 1, does not always produce convincing summaries. So, systems must balance the position bias with representations of the semantic content throughout the document. Alas, preliminary studies (Kedzie et al., 2018) suggest that even the most recent neural methods predominantly pick sentences from the lead, and that their content selection performance drops greatly when the position cues are withheld.

Lead-3: Bangladesh beat fellow World Cup quarter-finalists Pakistan by 79 runs in the first one-day international in Dhaka. Tamim Iqbal and Mushfiqur Rahim scored centuries as Bangladesh made 329 for six and Pakistan could only muster 250 in reply. Pakistan will have the chance to level the three-match series on Sunday when the second odi takes place in Mirpur.
Reference: Bangladesh beat fellow World Cup quarter-finalists Pakistan by 79 runs. Tamim Iqbal and Mushfiqur Rahim scored centuries for Bangladesh. Bangladesh made 329 for six and Pakistan could only muster 250 in reply. Pakistan will have the chance to level the three-match series on Sunday.
Lead-3: Standing up for what you believe. What does it cost you? What do you gain?
Reference: Indiana town's Memories Pizza is shut down after online threat. Its owners say they'd refuse to cater a same-sex couple's wedding.

Table 1: 'Lead' (first 3 sentences of source) can produce extremely faithful (top) to disastrously inaccurate (bottom) summaries. Gold standard summaries are also shown.

In this paper, we verify that sentence position and lead bias dominate the learning signal for state-of-the-art neural extractive summarizers in the news domain. We then present techniques to improve content selection in the face of this bias. The first technique makes use of 'unbiased data' created by permuting the order of sentences in

[*]Equal contribution.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 6019–6024, Hong Kong, China, November 3–7, 2019.

the training articles. We use this shuffled dataset for pre-training, followed by training on the original (unshuffled) articles. The second method introduces an auxiliary loss which encourages the model's scores for sentences to mimic an estimated score distribution over the sentences, the latter computed using ROUGE overlap with the gold standard. We implement these techniques for two recent reinforcement learning based systems, RNES (Wu and Hu, 2018) and BanditSum (Dong et al., 2018), and evaluate them on the CNN/Daily Mail dataset (Hermann et al., 2015).

We find that our auxiliary loss achieves significantly better ROUGE scores compared to the base systems, and that the improvement is even more pronounced when the true best sentences appear later in the article. On the other hand, the pre-training approach produces mixed results. We also confirm that when summary-worthy sentences appear late, there is a large performance discrepancy between the oracle summary and state-of-the-art summarizers, indicating that learning to balance lead bias with other features of news text is a noteworthy issue to tackle.

2 Related Work

Modern summarization methods for news are typically based on neural network-based sequence-to-sequence learning (Kalchbrenner et al., 2014; Kim, 2014; Chung et al., 2014; Yin and Pei, 2015; Cao et al., 2015; Cheng and Lapata, 2016; Nallapati et al., 2017; Narayan et al., 2018a; Zhou et al., 2018). In MLE-based training, extractive summarizers are trained with gradient ascent to maximize the likelihood of heuristically-generated ground-truth binary labels (Nallapati et al., 2017). Many MLE-based models do not perform as well as their reinforcement learning-based (RL) competitors that directly optimize ROUGE (Paulus et al., 2018; Narayan et al., 2018b; Dong et al., 2018; Wu and Hu, 2018). As RL-based models represent the state of the art for extractive summarization, we analyze them in this paper.

The closest work to ours is a recent study by Kedzie et al. (2018) which showed that MLE-based models learn a significant bias for selecting early sentences when trained on news articles as opposed to other domains. As much as 58% of selected summary sentences come directly from the lead. Moreover, when these models are trained on articles whose sentences are randomly shuffled, the performance drops considerably for news domain only. While this drop could be due to the destruction of position cues, it may also arise because the article's coherence and context were lost.

In this paper, we employ finer control on the distortion of sentence position, coherence, and context, and confirm that performance drops are mainly due to the lack of position cues. We also propose the first techniques to counter the effects of lead bias in neural extractive systems.

3 Base Models for Extractive Summarization

In supervised systems, given a document $D = \{s_1, ..., s_n\}$ with n sentences, a summary can be seen as set of binary labels $y_1, \ldots, y_n \in \{0, 1\}$, where $y_i = 1$ indicates that the i-th sentence is included in the summary.

We choose to experiment with two state-of-the-art RL-based extractive models: **RNES** (Wu and Hu, 2018) and **BanditSum** (Dong et al., 2018). Both employ an encoder-decoder structure, where the encoder extracts sentence features into fixed-dimensional vector representations $h_1, \ldots, h_n$, and a decoder produces the labels $y_1, \ldots, y_n$ based on these sentence representations. RNES uses a CNN+bi-GRU encoder, and BanditSum a hierarchical bi-LSTM. RNES's decoder is *auto-regressive*, meaning it predicts the current sentence's label based on decisions made on previous sentences; i.e., $y_t = f(D, h_t, y_{1:t-1})$. In BanditSum, there is no such dependence: it produces affinity scores for each sentence and the top scoring sentences are then selected.

4 Lead Bias of News Systems

First, we investigate the impact of sentence position on our models. We manipulate the **original** CNN/Daily Mail dataset to preserve sentence position information at different levels. In the **random** setting, sentences are shuffled randomly; in **reverse**, they are in reverse order; in **insert-lead** and **insert-lead3**, we insert an out-of-document sentence (chosen randomly from the corpus) as the first sentence or randomly as one of the first three sentences, respectively.

In Table 2, we show BanditSum's performance,[1] when trained and tested on the various datasets. All models (except random) perform

[1] We notice the same trends on RNES.

train setting \ test setting	original	random	reverse	insert-lead	insert-lead3	Mean	Std. Dev.
Lead-3 baseline	32.68	22.81	17.94	27.67	27.68	25.76	5.00
original	**33.85**	26.18	20.71	31.71	31.11	28.71	4.72
random	30.88	**29.70**	29.79	29.97	30.09	**30.09**	**0.42**
reverse	21.35	26.32	**33.59**	21.63	21.65	24.91	4.72
insert-lead	33.21	26.07	20.70	**33.41**	31.59	29.00	4.93
insert-lead3	32.29	25.57	20.22	32.92	**32.15**	28.63	4.98

Table 2: BanditSum's performance—calculated as the average between ROUGE-1,-2, and -L F1—on the validation set of the CNN/Daily Mail corpus. The sentence position information is perturbed at different levels, as explained in Section 4.

worse when tested on a mismatched data perturbation. Even when the distortion is at a single lead position in **insert-lead** and **insert-lead3**, the performance on the original data is significantly lower than when trained without the distortion. These results corroborate Kedzie et al. (2018)'s findings for RL-based systems. Interestingly, the **random** model has the best mean performance and the lowest variation indicating that completely removing the position bias may allow a model to focus on learning robust sentence semantics.

5 Learning to Counter Position Bias

We present two methods which encourage models to locate key phrases at diverse parts of the article.

5.1 Multi-Stage Training

This technique is inspired by the robust results from the **random** model in section 4. We implement a multi-stage training method for both BanditSum and RNES where in the first few epochs, we train on an 'unbiased' dataset where the sentences in every training document are randomly shuffled. We then fine-tune the models by training on the original training articles. The goal is to prime the model to learn sentence semantics independently of position, and then introduce the task of balancing semantics and positional cues.

5.2 ROUGE-based Auxiliary Loss

We observed that BanditSum tends to converge to a low-entropy policy, in the sense that the model's affinity scores are either 1 or 0 at the end of training. Furthermore, over 68% of its selections are from the three leading sentences of the source. Regularizing low-entropy policies can increase a model's propensity to explore potentially good states or stay close to a known good policy (Nachum et al., 2017; Galashov et al., 2019). We extend this idea to summarization by introducing a ROUGE-based loss which regularizes the model policy using an estimate of the value of individual sentences.

These sentence-level estimates are computed as a *distribution* P_R:

$$P_R(x = i) = \frac{r(s_i, \mathcal{G})}{\sum_{j=1}^{n} r(s_j, \mathcal{G})}, \tag{1}$$

where r is the average of ROUGE-1, -2 and -L F_1 scores between sentence s_i in the article and the reference summary $\mathcal{G}$. We would like the model's predictive distribution $P_\mathcal{M}$ to approximately match P_R. To compute $P_\mathcal{M}$, we normalize the predicted scores from a non-auto-regressive model. In an auto-regressive model such as RNES, the decision of including a sentence depends on those selected so far. So a straightforward KL objective is hard to implement, and we use this technique for BanditSum only.

Our auxiliary loss is defined as the KL divergence: $\mathcal{L}_{\text{KL}} = D_{\text{KL}}(P_R \parallel P_\mathcal{M})$. The update rule then becomes:

$$\theta^{(t+1)} = \theta^{(t)} + \alpha \left(\nabla \mathcal{L}_\mathcal{M}(\theta^{(t)}) + \beta \nabla \mathcal{L}_{\text{KL}}(\theta^{(t)}) \right) \tag{2}$$

where $\theta^{(t)}$ represents the model's parameters at time step t, $\mathcal{L}_\mathcal{M}$ is the original model's loss function, and β is a hyperparameter.

6 Experimental Setup

We use the CNN/Daily Mail dataset (Hermann et al., 2015) with the standard train/dev/test splits of 287,227/13,368/11,490. To avoid inconsistencies, we built on top of the author-provided implementations for BanditSum and our faithful reimplementation of RNES.

To reduce training time, we pre-compute and store the average of ROUGE-1, -2, and -L for every sentence triplet of each article, using a HDF5 table and PyTables (PyTables Developers Team, 2002-2019; The HDF Group, 1997-2019). This allows for a considerable increase in training speed.

Model	ROUGE			Overlp
	1	2	L	%
Lead-3	40.06	17.53	36.18	100.0
Oracle	56.53	32.65	53.12	27.24
Refresh	40.0	18.2	36.6	–
NeuSum	40.15	17.80	36.63	58.24
RNES	41.15	18.81	37.75	68.44
RNES+pretrain	41.29	18.85	37.79	68.22
BanditSum	41.68	18.78	38.00	69.87
B.Sum+pretrain	41.68	18.79	37.99	70.77
B.Sum+entropy	41.71	18.87	38.04	64.83
BanditSum+KL	**41.81***	**18.96***	**38.16***	65.13

Table 3: ROUGE scores for systems. 'Overlp' denotes the model's overlap in extraction choices with the lead-3 baseline. Scores significantly higher than BanditSum with $p < 0.001$ (bootstrap resampling test) are marked with *.

We limit the maximum number of sentences considered in an article to the first 100.

All the models were trained for 4 epochs. For the multi-stage training, we pretrain for 2 epochs, then train on the original articles for 2 epochs. We set the auxiliary loss hyperparameters $\alpha = 1e-4$ and $\beta = 0.0095$ in eq. 2 based on a grid search using the Tune library (Liaw et al., 2018).

We also train a baseline **entropy** model by replacing $\mathcal{L}_{\text{KL}}$ with the negated entropy of $P_{\mathcal{M}}$ in eq. 2. This loss penalizes low entropy, helping the model explore, but it is 'undirected' compared to our proposed method. We present the results of Lead-3 baseline (first 3 sentences), and two other competitive models—Refresh[2] (Narayan et al., 2018a) and NeuSum (Zhou et al., 2018).

Lastly, we include results from an *oracle* summarizer, computed as the triplet of source sentences with the highest average of ROUGE-1, -2 and -L scores against the abstractive gold standard.

7 Results and Discussion

Table 3 reports the F1 scores for ROUGE-1,-2 and -L (Lin, 2004). We use the *pyrouge*[3] wrapper library to evaluate the final models, while training with a faster Python-only implementation[4].

We test for significance between the baseline models and our proposed techniques using the bootstrap method. This method was first recommended for testing significance in ROUGE scores by Lin (2004), and has subsequently been advocated as an appropriate measure in works such as Dror et al. (2018) and Berg-Kirkpatrick et al. (2012).

[2]We are unable to evaluate this model on the lead overlap measure due to lack of access to the model outputs.

[3]www.github.com/bheinzerling/pyrouge

[4]www.github.com/Diego999/py-rouge

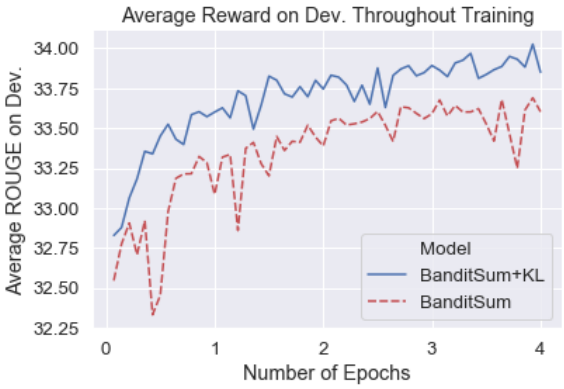

Figure 1: Training curves for BanditSum based models. Average ROUGE is the average of ROUGE-1, -2 and -L F1.

The simple entropy regularizer has a small but not significant improvement, and pretraining has a similar improvement only for RNES. But the auxiliary ROUGE loss significantly ($p < 0.001$) improves over BanditSum, obtaining an extra 0.15 ROUGE points on average. The last column reports the percentage of summary sentences which overlap with the lead. The auxiliary loss leads to a 4.7% absolute decrease in such selections compared to the base system, while also reaching a better ROUGE score. Figure 1 shows that the reward (average ROUGE-1,-2,-L) for the auxiliary loss model is consistently above the base.

We also examined the auxiliary loss model on documents where the summary is mostly comprised of lead sentences D_{early}, mostly sentences much later in the article D_{late}, and a dataset at the midway point, D_{med}. To create these sets, we rank test articles using the average index of its summary sentences in the source document. The 100 test articles with lowest average index are D_{early}, the 100 with highest value are D_{late} and the 100 closest to the median are D_{med}. In Table 4, we can see that the auxiliary loss model's improvements are even more amplified on D_{med} and D_{late}.

On the other hand, our pretraining results are mixed. We hope to employ more controlled multi-tasking methods (Kiperwasser and Ballesteros, 2018) in the future to deal with the issue.

The second line in Table 4 reports the oracle ROUGE scores of the best possible extractive summary. While all systems are quite close to the oracle on D_{early} they only reach half the performance on D_{late}. This gap indicates that our improvements only scratch the surface, but also that this problem is worthy and challenging to explore.

It is worth noting that we have attempted to build a single model which can summarize both

Model	D_{early}	D_{med}	D_{late}
Lead-3	46.17	30.90	20.18
Oracle	50.52	47.92	42.21
RNES	41.76	32.11	20.62
RNES+pretrain	41.66	32.38	20.64
BanditSum	43.10	32.65	21.63
BanditSum+entropy	41.96	32.59	22.12
BanditSum+KL	42.63	33.05	21.96

Table 4: Average ROUGE-1, -2 and -L F1 scores on D_{early}, and D_{med}, D_{late}. Each set contains 100 documents.

lead-biased articles and those whose information is spread throughout. Our aim was to encourage the model to explore useful regions as a way of learning better document semantics. But we hypothesize that our models can be further improved by learning to automatically predict when the lead paragraph suffices as a summary, and when the model should look further in the document.

8 Conclusion

In this paper, we have presented the first approaches for learning a summarization system by countering the strong effect of summary-worthy lead sentences. We demonstrate that recent summarization systems over-exploit the inherent lead bias present in news articles, to the detriment of their summarization capabilities. We explore two techniques aimed at learning to better balance positional cues with semantic ones. While our auxiliary loss method achieves significant improvement, we note that there is a large gap which better methods can hope to bridge in the future.

One approach, building on ours, is to examine other ways to combine loss signals (Finn et al., 2017), and to encourage exploration (Haarnoja et al., 2018). We will also carry out deeper study of the properties of D_{early} and D_{late} type documents and use them to inform new solutions. On cursory analysis, the most frequent terms in D_{early} tend to be about UK politics, while in D_{late} they are often related to British soccer.

Acknowledgments

This work is supported by the Natural Sciences and Engineering Research Council of Canada, the Institute for Data Valorisation (IVADO), Compute Canada, and the CIFAR Canada AI Chair program.

References

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.

Ronald Brandow, Karl Mitze, and Lisa F Rau. 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing & Management*, 31(5):675–685.

Ziqiang Cao, Furu Wei, Sujian Li, Wenjie Li, Ming Zhou, and Houfeng Wang. 2015. Learning summary prior representation for extractive summarization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning*.

Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. 2018. Banditsum: Extractive summarization as a contextual bandit. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3739–3748.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135, International Convention Centre, Sydney, Australia. PMLR.

Alexandre Galashov, Siddhant M Jayakumar, Leonard Hasenclever, Dhruva Tirumala, Jonathan Schwarz, Guillaume Desjardins, Wojciech M Czarnecki, Yee Whye Teh, Razvan Pascanu, and Nicolas Heess. 2019. Information asymmetry in kl-regularized rl. *International Conference on Learning Representations (ICLR)*.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with

a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1861–1870, Stockholmsmssan, Stockholm Sweden. PMLR.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

Kai Hong and Ani Nenkova. 2014. Improving the estimation of word importance for news multi-document summarization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 712–721.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Chris Kedzie, Kathleen McKeown, and Hal Daume III. 2018. Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Eliyahu Kiperwasser and Miguel Ballesteros. 2018. Scheduled multi-task learning: From syntax to translation. *Transactions of the Association of Computational Linguistics*, 6:225–240.

Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. 2018. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*.

Yang Liu. 2019. Fine-tune BERT for extractive summarization. *CoRR*, abs/1903.10318.

Ofir Nachum, Mohammad Norouzi, and Dale Schuurmans. 2017. Improving policy gradient by exploring under-appreciated rewards. In *International Conference on Learning Representations*, pages 2775–2785.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gùlçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL)*.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018a. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018b. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*.

Ani Nenkova. 2005. Automatic text summarization of newswire: Lessons learned from the document understanding conference. In *AAAI*, volume 5, pages 1436–1441.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations (ICLR)*.

PyTables Developers Team. 2002-2019. PyTables: Hierarchical datasets in Python.

Barry Schiffman, Ani Nenkova, and Kathleen McKeown. 2002. Experiments in multidocument summarization. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 52–58.

The HDF Group. 1997-2019. Hierarchical Data Format, version 5. Http://www.hdfgroup.org/HDF5/.

Yuxiang Wu and Baotian Hu. 2018. Learning to extract coherent summary via deep reinforcement learning. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*.

Wenpeng Yin and Yulong Pei. 2015. Optimizing sentence modeling and selection for document summarization. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1383–1389.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–663, Melbourne, Australia. Association for Computational Linguistics.