On the Abstractiveness of Neural Document Summarization

Fangfang Zhang[1*] **Jin-ge Yao**[1*] **Rui Yan**[1,2]
[1]Institute of Computer Science and Technology, Peking University, China
[2]Beijing Institute of Big Data Research, Beijing, China
{ffz, yaojinge, ruiyan}@pku.edu.cn

Abstract

Many modern neural document summarization systems based on encoder-decoder networks are designed to produce abstractive summaries. We attempted to verify the degree of abstractiveness of modern neural abstractive summarization systems by calculating overlaps in terms of various types of units. Upon the observation that many abstractive systems tend to be near-extractive in practice, we also implemented a pure copy system, which achieved comparable results as abstractive summarizers while being far more computationally efficient. These findings suggest the possibility for future efforts towards more efficient systems that could better utilize the vocabulary in the original document.

1 Introduction

Document summarization has been a hot research topic in natural language processing for long. When human writers summarize a document, they often edit its constituent sentences in order to succinctly capture the meaning of the document. For instance, Jing and McKeown (2000) observed that summary authors trimmed extraneous content, combined sentences, replaced phrases or clauses with more general or specific variants. The abstractive summaries thus involve sentences which deviate from those of the source document in structure or content.

On the contrary, automated summarization generally produces extractive summaries by selecting complete sentences from the source document (Nenkova et al., 2011) to ensure that the output is grammatical.

Recently, many modern neural summarization systems based on encoder-decoder networks have been proposed, aiming at producing abstractive summaries. These systems highly rely on the attention mechanism (Bahdanau et al., 2015) that focus on different parts of input during the decoding stage. Some also suggested to use a copying mechanism (Gulcehre et al., 2016; Gu et al., 2016) to directly copy words from input. This naturally brings us to a question: in how much degree are current neural document summarizers abstractive? In this work, we conduct such a study on the popularly-used CNN / DailyMail news corpora. By calculating various types of overlaps between summaries generated by neural abstractive summarizers and the original document, we verified that many systems are in fact heavily extracting text spans from input.

Recent studies found that automated methods can generate a wider range of summaries by extracting over sub-sentential units of meaning, such as elementary discourse units (EDUs), from the source documents rather than whole sentences (Li et al., 2016; Durrett et al., 2016). We built on a rather standard pointer-generator system to produce a summarizer that purely copies from input. Limited vocabulary size makes the new summarizer more computationally efficient, without loss of performance. The findings in this paper may hint future studies towards more efficient and more effective near-extractive systems, instead of a less important target of improving abstraction.

To summarize, in this paper we provide:

- A quantitative analysis on how abstractive current neural document summarizers are by calculating various types of content overlap with the input documents.
- A simple modification on the standard pointer-generator document summarizer to produce equally good near-extractive summaries while being more computationally efficient due to largely reduced vocabulary size.

[*] The first two authors contributed equally.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 785–790
Brussels, Belgium, October 31 - November 4, 2018.

2 Neural Abstractive Summarization

Recently end-to-end training with encoder-decoder neural networks (Sutskever et al., 2014) have achieved huge success in data sufficient sequence transduction tasks such as machine translation, which brings potential applications for summarization tasks, especially for abstractive settings. Earlier practice is mostly achieved on abstractive sentence summarization (Rush et al., 2015), which is essentially sentence simplification working on short text inputs. These neural sentence abstraction models are able to achieve good ROUGE (Lin and Hovy, 2003) scores on headline generation benchmarks, [1] but have not been proved to be useful for generating summaries with multiple sentences for full documents with longer contexts, which is the main focus of this study on document summarization.

One possible way for document-level neural summarization is to design hierarchical encoding to represent sentences and words at different levels. Related studies treat a document as a sequence of sentences and take sentence embeddings as input for a document-level RNN, while using a convolutional network or recurrent network to generate sentence vectors from original tokens (Cheng and Lapata, 2016). Meanwhile, the attention mechanism will become hierarchical as well (Nallapati et al., 2016). When decoding, sentence-level attention weights will be used as input for calculating word-level attention weights. Experimental results in previous work suggest that such schemes could be useful for extractive summarization when calculating sentence weights, but could only generate rather disappointing results for abstractive summaries.

It has been shown to be useful to incorporate the copying mechanism (Gulcehre et al., 2016; Gu et al., 2016; See et al., 2017) that allows a word to be generated by directly copying an input word rather than producing all words from the hidden state from scratch. Meanwhile, directly optimizing ROUGE via reinforcement learning has been shown to be more effective than optimizing reference likelihood (Paulus et al., 2018).

Recent work has achieved improvements by modeling attention based on more structured inter-sentence relationships such as graphs (Tan et al., 2017; Yasunaga et al., 2017). In practice, a severe issue of repetitive generation has been reported in other related work. It has been shown helpful to encourage diversity and novelty in calculating attention weights (Chen et al., 2016; Nema et al., 2017), or incorporating different modules with mutual communications to encode different paragraphs in the input document (Celikyilmaz et al., 2018). Another perspective is to promote better information coverage, such as pre-estimating term frequencies in the target summary (Suzuki and Nagata, 2017) or directly introducing a coverage loss between encoder states and decoder states (See et al., 2017).

Among the aforementioned related studies, a few proposed systems explicitly targeted at generating *abstractive* summaries for documents. However, these systems highly rely on the attention mechanism and/or copying mechanism that heavily depends on different part of input during the decoding stage. This naturally brings to a question on whether neural summarizers are indeed generating abstractive summaries after reading and digesting the input document, or they are just extracting subparts of the original document to perform near-extractive summaries.

3 Quantitative Analysis

3.1 Approaches

To verify whether current abstractive neural summarizers are just lazy generators that tend to copy original words and text spans, we computed the overlaps between the output summaries and the original article using the overlapping ratio measured by the following units: longest common subsequences (LCS), n-grams, and full sentences.

We studied a few representative systems that explicitly claim to have the ability to output *abstractive* summaries. The systems we compared in this work include:

- A basic sequence-to-sequence model with the attention mechanism.
- The pointer-generator system plus coverage mechanism presented by See et al. (2017).
- The graph-based attention system by Tan et al. (2017), aiming at better capturing salient information.
- The Distraction system by Chen et al. (2016), which attempt to distract attention.

[1] One caveat is that achieving high ROUGE scores on datasets with single references only is not an indication that the system is indeed generating good results.

- The deep reinforced model (Paulus et al., 2018) which combines intra-attention mechanism and reinforcement learning to target for better ROUGE scores and summary quality. to traverse between different content of a document so as to better grasp the overall meaning for summarization.

We also attempted to include some other popular abstractive document summarizers such as the SummaRuNNer system (Nallapati et al., 2017), but we failed to reach the authors to get their system outputs.

We collected experimental results from these systems on two large-scale corpora of CNN and DailyMail, which have been almost exclusively used in recent work on neural document summarization. The corpora were originally constructed in (Hermann et al., 2015) by collecting human generated abstractive highlights from the news stories. Just like almost all recent studies on neural summarization, the main conclusions might vary on other domains or even other news datasets.

3.2 Results

Table 1 displays the results of overlaps calculated for various system outputs over the original documents. Note that the authors of the Distraction system (Chen et al., 2016) did not conduct experiments on the Daily Mail subset of data, therefore only the results on CNN dataset are shown. We also include overlap results of manually-written reference summary highlights for comparison.

We can observe that the outputs from the pointer-generator systems (See et al., 2017) have the most amount of overlaps in terms of whole sentences, and most of the words or n-grams are in fact taken from the original document without further modifications or paraphrases. This observation is predictable since the system relies heavily on the pointer network module that directly copy from the input. The deep Reinforced model (Paulus et al., 2018), which relies on an intra-attention mechanism, also have rather high overlaps with the original document, suggesting that it is also an near-extractive system by some degree.

Other abstractive neural summarizers do not tend to directly copy full sentences, but do not generate words beyond the lexical choices used in the input document either, with considerably large overlaps of n-grams in general. A notable exception is the graph-based attention system where the overlap statistics are close to manually-written reference summaries. However, we manually checked a few samples and observed that the produced summaries tend to generate contents that do not conform to the information conveyed in the original documents. This is also verified in manual rating scores described later.

We conclude that currently many neural news summarizers which claimed to be abstractive tend to directly copy large spans of contents from the original documents, at least on the CNN / Daily Mail dataset which is the almost exclusively used benchmark in recent studies.

4 Near-Extractive Summarization

Now that we have observed large long-span overlaps between generated summaries and the original documents, it is natural to think about the following question: Do we really need to generate tokens from decoder states in a neural summarizer rather than just simply copying spans from input?

As previously mentioned, near-extractive summaries containing smaller text units from the input document have been shown sufficient for producing good summaries. On the other hand, generating words from a decoder state is based on time-consuming calculations of a softmax distribution, given that the vocabulary size is relatively large. Therefore, we would like to try abandoning decoder word generation, while just directly copying words from the input document instead.

4.1 Approach

We built a summarizer that only copy input words for outputs with trivial modifications upon the pointer-generator model (See et al., 2017). Specifically, it implements a sequence-to-sequence model that uses the soft attention distribution to produce an output sequence whose elements are all from the input sequence, similar to what a pointer network (Vinyals et al., 2015) does. We simply use the attention distribution as the final copy distribution, while the model does not generate words from the whole vocabulary using a softmax layer as in original recurrent networks. The training objective is to maximize the likelihood of words contained in reference summaries, similar to what has been used in the SummaRuNNer system for abstractive training (Nallapati et al., 2017). We keep using the same hyperparameters as in the original pointer-generator model.

	LCS	unigram	bigram	4-gram	sentence
Reference	75.0%	87.6%	49.0%	34.0%	3.5%
Seq2seq	87.4%	93.2%	76.0%	66.0%	10.8%
Pointer-generator	98.2%	99.8%	92.5%	93.0%	60.1%
Pointer-generator+coverage	98.8%	99.9%	96.1%	94.0%	70.0%
Reinforced (Paulus et al., 2018)	90.6%	95.8%	85.3%	80.2%	19.3%
Graph attention (Tan et al., 2017)	74.3%	82.3%	59.9%	42.1%	3.3%
Reference(CNN)	63.0%	75.2%	39.2%	25.9%	0.8%
Distraction (Chen et al., 2016) (CNN)	74.7%	94.0%	65.4%	38.7%	0.8%

Table 1: The overlap proportions between summaries and the original document

4.2 Experiments

We conducted experiments on the CNN / Daily Mail datasets and adopt the widely used ROUGE metrics (Lin and Hovy, 2003) for evaluation as previous work did.

Table 2 shows the ROUGE scores for the produced summaries. We can see that a pure copy system could produce equally or slightly better results in terms of word-matching metrics. The coverage mechanism introduced by See et al. (2017) is also effective for a pure copy system. As a side note, we verified the observation from See et al. (2017) that current neural summarizers cannot genuinely outperform a properly implemented LEAD baseline that simply takes the first three sentences from the original document, at least for the datasets used here that mainly consist of news describing events or activities.

	R-1	R-2	R-L
PtrGen	36.44	15.66	33.42
PtrGen + cov	39.53	17.28	36.38
Ptr	37.44	16.08	34.25
Ptr + cov	39.74	**17.31**	**36.53**
Lead3	**39.98**	17.25	36.20

Table 2: ROUGE scores on CNN/Daily Mail

We also conducted human evaluation on outputs for a sample of 30 documents in the common subpart of the system inputs. We asked three raters to evaluate on the following metrics in a summary using 1-5 scoring scheme (5 is the best, and rational numbers were allowed if raters felt uncertain over some cases): *informativeness* (INF), *relevance* (REL), *fluency* (FLU) and *coherence* (COH), as used by Grusky et al. (2018). The results are listed in Table 3. We find the pure copy system performs similarly to the pointer-generator. However, we can also observe a rough trend that: the more abstractive a system is, the higher the chance of generating irrelevant or grammatically worse content. Such observation is consistent with manual evaluation results conducted by another study towards more abstraction (Kryściński et al., 2018), in parallel to our work. This is a signal that other than pursuing for heavily abstractiveness, we could also spend more efforts on directly identifying and extracting useful pieces from the input, in order to get more controlled and more useful summaries with better quality beyond the heavily biased ROUGE metrics (Chaganty et al., 2018).

	INF	REL	FLU	COH
Lead3	2.97	3.20	4.10	3.33
Pure copy	3.03	3.37	3.87	3.14
PtrGen	3.01	3.44	3.65	3.11
Reinforced	3.18	3.17	3.57	2.97
Distraction	2.44	2.88	3.37	2.70
GraphAtt	2.75	2.87	2.47	2.14
Reference	3.71	4.32	4.54	4.12

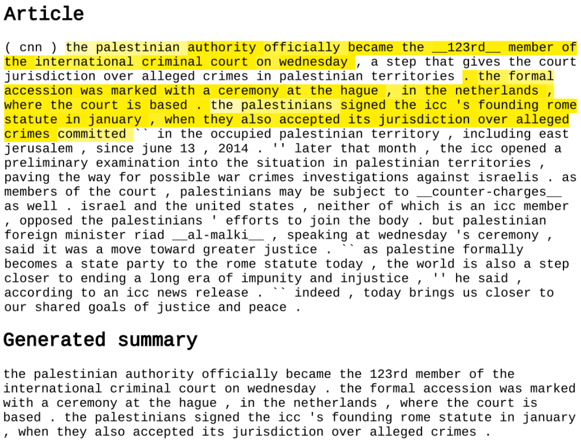
Table 3: Human ratings

A larger merit for implementing a pure copy system is that it is more computationally effective than abstractive summarizers that generate words in a large vocabulary from decoder states. Table 4 lists the speed for decoding as well as the memory costs in training for the pure copy system, compared with the original pointer-generator system. The numbers are averaged results from multiple runs on the same computing environment of Tesla M40 GPU. We can see that the speed doubles from a pure copy summarizer, while the GPU memory cost is reduced to around a quarter.

We visualize the copying probabilities (attention weights) for an example summary along with the original document in Figure 1. We can observe that the pure copy system tends to attend on con-

	Pure copy	PtrGen
Decoding Speed	0.87 step/s	0.44 step/s
GPU Usage	2216MB	8368MB
Memory Cost	2.67GB	3.30GB

Table 4: Comparison of computational costs



```
Article

( cnn ) the palestinian authority officially became the __123rd__ member of
the international criminal court on wednesday , a step that gives the court
jurisdiction over alleged crimes in palestinian territories . the formal
accession was marked with a ceremony at the hague , in the netherlands ,
where the court is based . the palestinians signed the icc 's founding rome
statute in january , when they also accepted its jurisdiction over alleged
crimes committed `` in the occupied palestinian territory , including east
jerusalem , since june 13 , 2014 . '' later that month , the icc opened a
preliminary examination into the situation in palestinian territories ,
paving the way for possible war crimes investigations against israelis . as
members of the court , palestinians may be subject to __counter-charges__
as well . israel and the united states , neither of which is an icc member
, opposed the palestinians ' efforts to join the body . but palestinian
foreign minister riad __al-malki__ , speaking at wednesday 's ceremony ,
said it was a move toward greater justice . `` as palestine formally
becomes a state party to the rome statute today , the world is also a step
closer to ending a long era of impunity and injustice , '' he said ,
according to an icc news release . `` indeed , today brings us closer to
our shared goals of justice and peace .

Generated summary

the palestinian authority officially became the 123rd member of the
international criminal court on wednesday . the formal accession was marked
with a ceremony at the hague , in the netherlands , where the court is
based . the palestinians signed the icc 's founding rome statute in january
, when they also accepted its jurisdiction over alleged crimes .
```

Figure 1: Visualization of copy probabilities

tinuous spans of input to form sentences.

5 Conclusion

In this work we attempted to quantify the abstractiveness of modern neural abstractive summarization systems by calculating overlaps of various units. Inspired by the observation that many systems tend to be near-extractive, we also implemented a pure copy system and achieved comparable performance while being far more efficient. Giving the observations that the abstractive summaries produced by current systems have lower quality than extractive summaries, our study should give hints for focusing on better extractions from the input, rather than deliberately pursuing for more abstraction but losing real quality beyond automatic metrics.

Acknowledgments

We would like to thank Abigail See for her open-sourced implementation of the pointer-generator summarizer, and Jiwei Tan, Qian Chen, Romain Paulus for providing the outputs from their systems. We also thank the anonymous reviewers for helpful comments. This work was supported by the National Key Research and Development Program of China (No. 2017YFC0804001), the National Science Foundation of China (NSFC No. 61876196 and No. 61672058). Rui Yan was sponsored by CCF-Tencent Open Research Fund and Microsoft Research Asia (MSRA) Collaborative Research Program. Rui Yan is the corresponding author of this paper.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *International Conference on Learning Representations (ICLR)*.

Asli Celikyilmaz, Antoine Bosselut, Xiaodong He, and Yejin Choi. 2018. Deep communicating agents for abstractive summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1662–1675, New Orleans, Louisiana. Association for Computational Linguistics.

Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debasing automatic metrics in natural language evaluation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Melbourne, Australia. Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, and Hui Jiang. 2016. Distraction-based neural networks for document summarization. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2754–2760.

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 (Volume 1: Long Papers)*, pages 484–494, Berlin, Germany. Association for Computational Linguistics.

Greg Durrett, Taylor Berg-Kirkpatrick, and Dan Klein. 2016. Learning-based single-document summarization with compression and anaphoricity constraints. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1998–2008, Berlin, Germany. Association for Computational Linguistics.

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)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany. Association for Computational Linguistics.

Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. 2016. Pointing the unknown words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 140–149, Berlin, Germany. Association for Computational Linguistics.

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 28*, pages 1693–1701.

Hongyan Jing and Kathleen R McKeown. 2000. Cut and paste based text summarization. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 178–185. Association for Computational Linguistics.

Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher. 2018. Improving abstraction in text summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Junyi Jessy Li, Kapil Thadani, and Amanda Stent. 2016. The role of discourse units in near-extractive summarization. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 137–147, Los Angeles. Association for Computational Linguistics.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram cooccurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 71–78. Association for Computational Linguistics.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents. In *AAAI Conference on Artificial Intelligence*, pages 3075–3081.

Ramesh Nallapati, Bowen Zhou, Ca glar Gulcehre, 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*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Preksha Nema, Mitesh M. Khapra, Anirban Laha, and Balaraman Ravindran. 2017. Diversity driven attention model for query-based abstractive summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1063–1072, Vancouver, Canada. Association for Computational Linguistics.

Ani Nenkova, Kathleen McKeown, et al. 2011. Automatic summarization. *Foundations and Trends® in Information Retrieval*, 5(2–3):103–233.

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

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. 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. Association for Computational Linguistics.

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)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, pages 3104–3112.

Jun Suzuki and Masaaki Nagata. 2017. Cutting-off redundant repeating generations for neural abstractive summarization. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 291–297, Valencia, Spain. Association for Computational Linguistics.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. Abstractive document summarization with a graph-based attentional neural model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1171–1181, Vancouver, Canada. Association for Computational Linguistics.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems 28*, pages 2692–2700.

Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir Radev. 2017. Graph-based neural multi-document summarization. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 452–462, Vancouver, Canada. Association for Computational Linguistics.