BIGPATENT: A Large-Scale Dataset for Abstractive and Coherent Summarization

Eva Sharma[1], **Chen Li**[2], and **Lu Wang**[1]
[1]Khoury College of Computer Sciences, Northeastern University
[2]Tencent AI Lab
[1]evasharma@ccs.neu.edu, luwang@ccs.neu.edu
[2]ailabchenli@tencent.com

Abstract

Most existing text summarization datasets are compiled from the news domain, where summaries have a flattened discourse structure. In such datasets, summary-worthy content often appears in the beginning of input articles. Moreover, large segments from input articles are present verbatim in their respective summaries. These issues impede the learning and evaluation of systems that can understand an article's global content structure as well as produce abstractive summaries with high compression ratio. In this work, we present a novel dataset, BIGPATENT, consisting of 1.3 million records of U.S. patent documents along with human written abstractive summaries. Compared to existing summarization datasets, BIGPATENT has the following properties: i) summaries contain a richer discourse structure with more recurring entities, ii) salient content is evenly distributed in the input, and iii) lesser and shorter extractive fragments are present in the summaries. Finally, we train and evaluate baselines and popular learning models on BIGPATENT to shed light on new challenges and motivate future directions for summarization research.

1 Introduction

There has been a growing interest in building neural abstractive summarization systems (See et al., 2017; Paulus et al., 2017; Gehrmann et al., 2018a), which requires large-scale datasets with high quality summaries. A number of summarization datasets have been explored so far (Sandhaus, 2008; Napoles et al., 2012; Hermann et al., 2015; Grusky et al., 2018). However, as most of them are acquired from news articles, they share specific characteristics that limit current state-of-the-art models by making them more extractive rather than allowing them to understand input content and generate well-formed informative summaries.

Sample CNN/Daily Mail News Summary
An *explosion* rocks a chemical plant in China's southeastern Fujian province for the second time in two years. Six were injured after the *explosion* and are being hospitalized. The *explosion* was triggered by an oil leak, though local media has not reported any toxic chemical spills.

Sample BIGPATENT Summary
A *shoelace cover* incorporating an interchangeable fashion panel for covering the *shoelaces* of a *gym shoe*. The *shoelace cover* is secured to the shoe by a number of *straps* threaded through slots in the *shoelace cover*. These *straps* secured to each side of the *gym shoe* include a loop and hook material such that the *straps* can be disengaged and the *shoelace cover* can be drawn back to expose the *shoelaces*. . .

Figure 1: Sample summaries from CNN/Daily Mail and BIGPATENT. Extractive fragments reused from input are underlined. Repeated entities indicating discourse structure are highlighted in respective colors.

Specifically, in these datasets, the summaries are flattened narratives with a simpler discourse structure, e.g., entities are rarely repeated as illustrated by the news summary in Fig. 1. Moreover, these summaries usually contain long fragments of text directly extracted from the input. Finally, the summary-worthy salient content is mostly present in the beginning of the input articles.

We introduce BIGPATENT[1], a new large-scale summarization dataset consisting of 1.3 million patent documents with human-written abstractive summaries. BIGPATENT addresses the aforementioned issues, thus guiding summarization research to better understand the input's global structure and generate summaries with a more complex and coherent discourse structure. The key features of BIGPATENT are: i) summaries exhibit a richer discourse structure with entities re-

[1]BIGPATENT dataset is available to download online at evasharma.github.io/bigpatent.

curring in multiple subsequent sentences as shown in Fig. 1, ii) salient content is evenly distributed in the document, and iii) summaries are considerably more abstractive while reusing fewer and shorter phrases from the input.

To further illustrate the challenges in text summarization, we benchmark BIGPATENT with baselines and popular summarization models, and compare with the results on existing large-scale news datasets. We find that many models yield noticeably lower ROUGE scores on BIGPATENT than on the news datasets, suggesting a need for developing more advanced models to address the new challenges presented by BIGPATENT. Moreover, while existing neural abstractive models produce more abstractive summaries on BIGPATENT, they tend to repeat irrelevant discourse entities excessively, and often fabricate information.

These observations demonstrate the importance of BIGPATENT in steering future research in text summarization towards global content modeling, semantic understanding of entities and relations, and discourse-aware text planning to build abstractive and coherent summarization systems.

2 Related Work

Recent advances in abstractive summarization show promising results in generating fluent and informative summaries (Rush et al., 2015; Nallapati et al., 2016; Tan et al., 2017; Paulus et al., 2017). However, these summaries often contain fabricated and repeated content (Cao et al., 2018). Fan et al. (2018) show that, for content selection, existing models rely on positional information and can be easily fooled by adversarial content present in the input. This underpins the need for global content modeling and semantic understanding of the input, along with discourse-aware text planning to yield a well-formed summary (McKeown, 1985; Barzilay and Lapata, 2008).

Several datasets have been used to aid the development of text summarization models. These datasets are predominantly from the news domain and have several drawbacks such as limited training data (Document Understanding Conference[2]), shorter summaries (Gigaword (Napoles et al., 2012), XSum (Narayan et al., 2018), and Newsroom (Grusky et al., 2018)), and near-extractive summaries (CNN / Daily Mail dataset (Hermann et al., 2015)). Moreover, due to the nature of news reporting, summary-worthy content is non-uniformly distributed within each article. ArXiv and PubMed datasets (Cohan et al., 2018), which are collected from scientific repositories, are limited in size and have longer yet extractive summaries. Thus, existing datasets either lack crucial structural properties or are limited in size for learning robust deep learning methods. To address these issues, we present a new dataset, BIGPATENT, which guides research towards building more abstractive summarization systems with global content understanding.

[2]https://duc.nist.gov/

Dataset	# Doc	Comp. ratio	Dens.	Summary # word	Summary # sent	Doc # word
CNN/DM	312,085	13.0	3.8	55.6	3.8	789.9
NYT	654,788	12.0	2.4	44.9	2.0	795.9
NEWSROOM	1,212,726	43.0	9.5	30.4	1.4	750.9
XSUM	226,711	18.8	1.2	23.3	1.0	431.1
ARXIV	215,913	39.8	3.8	292.8	9.6	6,913.8
PUBMED	133,215	16.2	5.8	214.4	6.9	3,224.4
BIGPATENT	1,341,362	36.4	2.4	116.5	3.5	3,572.8

Table 1: Statistics of BIGPATENT and other summarization datasets. # Doc: raw number of documents in each dataset. For all other columns, mean values are reported over all documents. BIGPATENT has a lower extractive fragment density (Dens.) and a higher compression ratio (Comp. ratio).

3 BIGPATENT Dataset

We present BIGPATENT, a dataset consisting of 1.3 million U.S. patent documents collected from Google Patents Public Datasets using BigQuery (Google, 2018)[3]. It contains patents filed after 1971 across nine different technological areas. We use each patent's abstract as the gold-standard summary and its description as the input.[4] Additional details for the dataset, including the preprocessing steps, are in Appendix A.1.

Table 1 lists statistics, including compression ratio and extractive fragment density, for BIGPATENT and some commonly-used summarization corpora. Compression ratio is the ratio of the number of words in a document and its summary, whereas density is the average length of the ex-

[3]Released and maintained by IFI CLAIMS Patent Services and Google, and licensed under *Creative Commons Attribution 4.0 International License*.

[4]The summarization task studied using BIGPATENT is notably different from traditional patent summarization task where patent claims are summarized into a more readable format (Cinciruk, 2015).

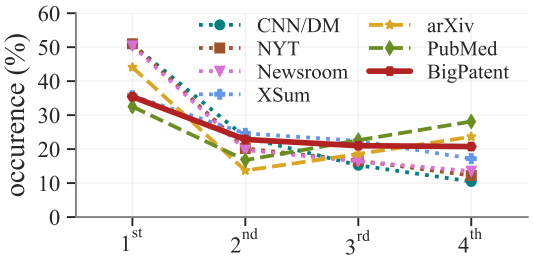

Figure 2: % of salient unigrams present in the N^{th} segments of the input.

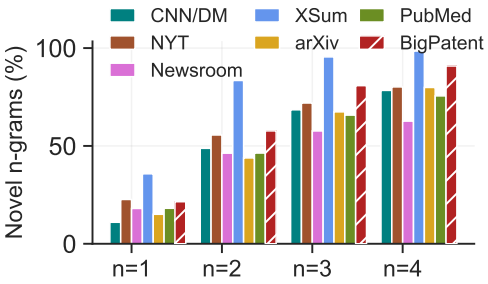

Figure 3: % of novel n-grams in the summaries.

tractive fragment[5] to which each word in the summary belongs (Grusky et al., 2018). Among existing datasets, CNN/DM (Hermann et al., 2015), NYT (Napoles et al., 2012), NEWSROOM (released) (Grusky et al., 2018) and XSUM (Narayan et al., 2018) are news datasets, while ARXIV and PUBMED (Cohan et al., 2018) contain scientific articles. Notably, BIGPATENT is significantly larger with longer inputs and summaries.

4 Dataset Characterization

4.1 Salient Content Distribution

Inferring the distribution of salient content in the input is critical to content selection of summarization models. While prior work uses probabilistic topic models (Barzilay and Lee, 2004; Haghighi and Vanderwende, 2009) or relies on classifiers trained with sophisticated features (Yang et al., 2017), we focus on salient words and their occurrences in the input.

We consider all unigrams, except stopwords, in a summary as *salient* words for the respective document. We divide each document into four equal segments and measure the percentage of unique salient words in each segment. Formally, let U be a function that returns all unique unigrams (except stopwords) for a given text. Then, $U(d^i)$ denotes the unique unigrams in the i^{th} segment of a document d, and $U(y)$ denotes the unique unigrams in the corresponding summary y. The percentage of salient unigrams in the i^{th} segment of a document is calculated as:

$$\frac{|(U(d^i) \cap U(y))| \times 100}{|U(y)|}\%$$

Fig. 2 shows that BIGPATENT has a fairly even distribution of salient words in all segments of the input. Only 6% more salient words are observed in the 1^{st} segment than in other segments. In contrast, for CNN/DM, NYT and Newsroom, approximately 50% of the salient words are present in the 1^{st} segment, and the proportion drops monotonically to 10% in the 4^{th} segment. This indicates that most salient content is present in the beginning of news articles in these datasets. For XSum, another news dataset, although the trend in the first three segments is similar to BIGPATENT, the percentage of novel unigrams in the last segment drops by 5% compared to 0.2% for BIGPATENT.

For scientific articles (arXiv and PubMed), where content is organized into sections, there is a clear drop in the 2^{nd} segment where related work is often discussed, with most salient information being present in the first (introduction) and last (conclusion) sections. Whereas in BIGPATENT, since each embodiment of a patent's invention is sequentially described in its document, it has a more uniform distribution of salient content.

Next, we probe how far one needs to read from the input's start to cover the salient words (only those present in input) from the summary. About 63% of the sentences from the input are required to construct full summaries for CNN/DM, 57% for XSum, 53% for NYT, and 29% for Newsroom. Whereas in the case of BIGPATENT, **80**% of the input is required. The aforementioned observations signify the need of global content modeling to achieve good performance on BIGPATENT.

4.2 Summary Abstractiveness and Coherence

Summary n-gram Novelty. Following prior work (See et al., 2017; Chen and Bansal, 2018), we compute abstractiveness as the fraction of novel n-grams in the summaries that are absent from the input. As shown in Fig. 3, XSum comprises of notably shorter but more abstractive summaries. Besides that, BIGPATENT reports the sec-

[5]Extractive fragments are the set of shared sequences of tokens in the document and summary.

	$t=1$	$t=2$	$t=3$	$t\geq 3$
CNN/DM	95.7%	3.9%	0.4%	0.1%
NYT	97.6%	2.1%	0.3%	0.1%
NEWSROOM	98.9%	1.0%	0.1%	0.02%
ARXIV	89.5%	7.9%	1.7%	0.9%
PUBMED	86.1%	9.3%	2.7%	2.0%
BIGPATENT	75.9%	15.1%	5.1%	3.9%

Table 2: % of entities occurring t times in summaries.

	Ent. Chain Length (In %)				Ent. Recurrence at		
Datasets	$l=1$	$l=2$	$l=3$	$l>3$	$t+1$	$t+2$	$\geq t+3$
CNN/DM	97.7	2.1	0.2	0.02	0.3	0.2	0.2
NYT	98.7	1.2	0.1	0.01	0.4	0.2	0.1
NEWSROOM	99.6	0.4	0.02	0.002	0.2	0.1	0.1
ARXIV	95.6	3.8	0.5	0.1	1.6	1.0	3.8
PUBMED	93.9	4.9	0.9	0.3	2.0	1.1	2.1
BIGPATENT	85.9	11.1	2.3	0.7	2.4	1.1	1.2

Table 3: Left: % of entities of chain length l. Right: Avg. number of entities that appear at the t^{th} summary sentence and recur in a later sentence.

ond highest percentage of novel n-grams, for $n \in \{2,3,4\}$. Significantly higher novelty scores for trigram and 4-gram indicate that BIGPATENT has fewer and shorter extractive fragments, compared to others (except for XSum, a smaller dataset). This further corroborates the fact that BIGPATENT has the lowest extractive fragment density (as shown in Table 1) and contains longer summaries.

Coherence Analysis via Entity Distribution. To study the discourse structure of summaries, we analyze the distribution of entities that are indicative of coherence (Grosz et al., 1995; Strube and Hahn, 1999). To identify these entities, we extract non-recursive noun phrases (regex `NP` $\rightarrow$ `ADJ*[NN]+`) using NLTK (Loper and Bird, 2002). Finally, we use the entity-grid representation by Barzilay and Lapata (2008) and their coreference resolution rules to capture the entity distribution across summary sentences. In this work, we do not distinguish entities' grammar roles, and leave that for future study.

On average, there are 6.7, 10.9, 12.4 and 18.5 unique entities in the summaries for Newsroom, NYT, CNN/DM and BIGPATENT, respectively[6]. PUBMED and ARXIV reported higher number of unique entities in summaries (39.0 and 48.1 respectively) since their summaries are considerably longer (Table 1). Table 2 shows that 24.1% of entities recur in BIGPATENT summaries, which is higher than that on other datasets, indicating more complex discourse structures in its summaries. To understand local coherence in summaries, we measure the longest chain formed across sentences by each entity, denoted as l. Table 3 shows that 11.1% of the entities in BIGPATENT appear in two consecutive sentences, which is again higher than that of any other dataset. The presence of longer entity chains in the BIGPATENT summaries suggests its higher sentence-to-sentence relatedness than the news summaries.

Finally, we examine the entity recurrence pattern which captures how many entities, first occurring in the t^{th} sentence, are repeated in subsequent $(t+i^{th})$ sentences. Table 3 (right) shows that, on average, 2.3 entities in BIGPATENT summaries recur in later sentences (summing up the numbers for t+2 and after). The corresponding recurring frequency for news dataset such as CNN/DM is only 0.4. Though PUBMED and ARXIV report higher number of recurrence, their patterns are different, i.e., entities often recur after three sentences. These observations imply a good combination of local and global coherence in BIGPATENT.

5 Experiments and Analyses

We evaluate BIGPATENT with popular summarization systems and compare with well-known datasets such as CNN/DM and NYT. For baseline, we use LEAD-3, which selects the first three sentences from the input as the summary. We consider two oracles: i) ORACLEFRAG builds summary using all the longest fragments reused from input in the gold-summary (Grusky et al., 2018), and ii) ORACLEEXT selects globally optimal combination of three sentences from the input that gets the highest ROUGE-1 F1 score. Next, we consider three unsupervised extractive systems: TEXTRANK (Mihalcea and Tarau, 2004), LEXRANK (Erkan and Radev, 2004), and SUMBASIC (Nenkova and Vanderwende, 2005). We also adopt RNN-EXT RL (Chen and Bansal, 2018), a SEQ2SEQ model that selects three salient sentences to construct the summary using reinforcement learning. Finally, we train four abstractive systems: SEQ2SEQ with attention, Pointer-Generator (POINTGEN) and a version with coverage mechanism (POINTGEN + COV) (See et al., 2017), and SENTREWRITING (Chen and Bansal, 2018). Experimental setups and model parameters are described in Appendix A.2.

Table 4 reports F1 scores of ROUGE-1, 2,

[6]We exclude XSum as its summaries are all one-sentence.

Models	CNN/DM			NYT			BIGPATENT		
	R-1	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
LEAD-3	40.23	17.52	36.34	32.93	17.69	29.58	31.27	8.75	26.18
ORACLEFRAG (Grusky et al., 2018)	93.36	83.19	93.36	88.15	74.74	88.15	91.85	78.66	91.85
ORACLEEXT	49.35	27.96	46.24	42.62	26.39	39.50	43.56	16.91	36.52
TEXTRANK (Mihalcea and Tarau, 2004)	37.72	15.59	33.81	28.57	14.29	23.79	35.99	11.14	29.60
LEXRANK (Erkan and Radev, 2004)	33.96	11.79	30.17	27.32	11.93	23.75	35.57	10.47	29.03
SUMBASIC (Nenkova and Vanderwende, 2005)	31.72	9.60	28.58	23.16	7.18	20.06	27.44	7.08	23.66
RNN-EXT RL (Chen and Bansal, 2018)	**41.47**	**18.72**	**37.76**	39.15	22.60	34.99	34.63	10.62	29.43
SEQ2SEQ (Sutskever et al., 2014)	31.10	11.54	28.56	41.57	26.89	38.17	28.74	7.87	24.66
POINTGEN (See et al., 2017)	36.15	15.11	33.22	43.49	28.70	39.66	30.59	10.01	25.65
POINTGEN+COV (See et al., 2017)	39.23	17.09	36.03	**45.13**	**30.13**	39.67	33.14	11.63	28.55
SENTREWRITING (Chen and Bansal, 2018)	40.04	17.61	37.59	44.77	29.10	**41.55**	**37.12**	**11.87**	**32.45**

Table 4: ROUGE scores on three large datasets. The best results for non-baseline systems are in bold. Except for SentRewriting on CNN/DM and NYT, for all abstractive models, we truncate input and summaries at 400 and 100.

Models	**% Novel n-grams**		**% Entities Occurring m Times**			
	$n=1$	$n=2$	$m=1$	$m=2$	$m=3$	$m>3$
GOLD	21.5%	57.7%	75.5%	15.2%	5.2%	4.0%
SEQ2SEQ	**18.6%**	**52.0%**	51.4%	19.4%	6.7%	22.6%
POINTGEN + COV	9.7%	33.9%	**82.7%**	**13.8%**	**2.4%**	**1.2%**
SENTREWRITING	11.5%	44.9%	69.5%	17.3%	6.6%	6.6%

Table 5: % of novel n-grams (highest % are highlighted), and % of entities occurring m times in generated summaries of BIGPATENT. POINTGEN+COV repeats entities less often than humans do.

and L (Lin and Hovy, 2003) for all models. For BIGPATENT, almost all models outperform the LEAD-3 baseline due to the more uniform distribution of salient content in BIGPATENT's input articles. Among extractive models, TEXTRANK and LEXRANK outperform RNN-EXT RL which was trained on only the first 400 words of the input, again suggesting the need for neural models to efficiently handle longer input. Finally, SENTREWRITING, a reinforcement learning model with ROUGE as reward, achieves the best performance on BIGPATENT.

Table 5 presents the percentage of novel n-grams in the generated summaries. Although the novel content in the generated summaries (for both unigrams and bigrams) is comparable to that of GOLD, we observe repeated instances of fabricated or irrelevant information. For example, *"the upper portion is configured to receive the upper portion of the sole portion"*, part of SEQ2SEQ generated summary has irrelevant repetitions compared to the human summary as in Fig. 1. This suggests the lack of semantic understanding and control for generation in existing neural models.

Table 5 also shows the entity distribution (§4.2) in the generated summaries for BIGPATENT. We find that neural abstractive models (except POINTGEN+COV) tend to repeat entities more often than humans do. For GOLD, only 5.2% and 4.0% of entities are mentioned thrice or more, compared to 6.7% and 22.6% for SEQ2SEQ. POINTGEN+COV, which employs coverage mechanism to explicitly penalize repetition, generates significantly *fewer* entity repetitions. These findings indicate that current models failt to learn the entity distribution pattern, suggesting a lack of understanding of entity roles (e.g., their importance) and discourse-level text planning.

6 Conclusion

We present the BIGPATENT dataset with human-written abstractive summaries containing fewer and shorter extractive phrases, and a richer discourse structure compared to existing datasets. Salient content from the BIGPATENT summaries is more evenly distributed in the input. BIGPATENT can enable future research to build robust systems that generate abstractive and coherent summaries.

Acknowledgements

This research is supported in part by National Science Foundation through Grants IIS-1566382 and IIS-1813341, and by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract # FA8650-17-C-9116. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. We also thank the anonymous reviewers for their constructive suggestions.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Federico Barrios, Federico López, Luis Argerich, and Rosa Wachenchauzer. 2016. Variations of the similarity function of textrank for automated summarization. *arXiv preprint arXiv:1602.03606*.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1).

Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686. Association for Computational Linguistics.

David Cinciruk. 2015. Patent summarization and paraphrasing. http://www.ece.drexel.edu/walsh/David_PatentSummarization.pdf.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621. Association for Computational Linguistics.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Lisa Fan, Dong Yu, and Lu Wang. 2018. Robust neural abstractive summarization systems and evaluation against adversarial information. In *Workshop on Interpretability and Robustness in Audio, Speech, and Language (IRASL)*. Neural Information Processing Systems.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018a. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109. Association for Computational Linguistics.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018b. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109.

Google. 2018. Google patents public datasets: connecting public, paid, and private patent data. https://console.cloud.google.com/marketplace/details/google_patents_public_datasets/google-patents-public-data?_ga=2.148226999.-1648178590.1534442735&pli=1. Accessed: 2018-08-30.

Barbara J. Grosz, Scott Weinstein, and Aravind K. Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2).

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. Association for Computational Linguistics.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370. 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*, pages 1693–1701.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*.

Kathleen R McKeown. 1985. Discourse strategies for generating natural-language text. *Artificial Intelligence*, 27(1):1–41.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar 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. Association for Computational Linguistics.

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

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. 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. Association for Computational Linguistics.

Ani Nenkova and Lucy Vanderwende. 2005. The impact of frequency on summarization. *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005*, 101.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

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.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

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. Association for Computational Linguistics.

Michael Strube and Udo Hahn. 1999. Functional centering grounding referential coherence in information structure. *Computational Linguistics*, 25(3).

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

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. Association for Computational Linguistics.

USPTO. 2013. Cooperative patent classification scheme. `https://www.uspto.gov/web/patents/classification/cpc/html/cpc.html`. Accessed: 2018-08-30.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yinfei Yang, Forrest Bao, and Ani Nenkova. 2017. Detecting (un)important content for single-document news summarization. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 707–712. Association for Computational Linguistics.

A Appendices

A.1 Dataset Details

BIGPATENT, a novel large-scale summarization dataset of 1.3 million US Patent documents, is collected from Google Patents Public Datasets using BigQuery (Google, 2018). Google has indexed more than 87 million patents with full text from 17 different patent offices so far. We only consider patent documents from United States Patent and Trademark Office (USPTO) filed in English language after 1971 in order to get considerably more consistent writing and formatting style to facilitate easier parsing of the text.

Each US patent application is filed under a Cooperative Patent Classification (CPC) code (USPTO, 2013) that provides a hierarchical system of language independent symbols for the classification of patents according to the different areas of technology to which they pertain. There are nine such classification categories: A (Human Necessities), B (Performing Operations; Transporting), C (Chemistry; Metallurgy), D (Textiles; Paper), E (Fixed Constructions), F (Mechanical

CPC code	# Doc	Comp. ratio	Dens.	Summary		Doc
				# word	# sent	# word
A	193,483	39.5	2.3	109.5	3.4	3,520.7
B	179,467	28.1	2.3	116.6	3.4	2,900.4
C	112,269	71.3	2.6	97.9	2.6	5,278.4
D	11,294	30.1	2.3	113.0	3.2	2,892.1
E	38,271	26.9	2.2	117.2	3.7	2,814.3
F	95,076	26.0	2.3	116.7	3.5	2,737.8
G	287,706	35.9	2.4	123.7	3.6	3,924.1
H	285,577	32.7	2.4	121.1	3.6	3,531.4
Y	138,219	33.5	2.3	116.3	3.5	3,328.0

Table 6: Statistics for 9 CPC codes in BIGPATENT.

Engineering; Lightning; Heating; Weapons; Blasting), G (Physics), H (Electricity), and Y (General tagging of new or cross-sectional technology). Table 6 summarizes the statistics for BIGPATENT across all nine categories.

From the full public dataset, for each patent record, we retained its title, authors, abstract, claims of the invention and the description text. Abstract of the patent, which is generally written by the inventors after the patent application is approved, was considered as the gold-standard summary of the patent. Description text of the patent contains several other fields such as background of the invention covering previously published related inventions, description of figures, and detailed description of the current invention. For the summarization task, we considered the detailed description of each patent as the input.

We tokenized the articles and summaries using Natural Language Toolkit (NLTK) (Bird et al., 2009). Since there was a large variation in size of summary and input texts, we removed patent records with compression ratio less than 5 and higher than 500. Further, we only kept records with summary length between 10 and 2,500 words, and input length of at least 150 and at most 80,000. Next, to focus on the abstractive summary-input pairs, we removed the records whose percentage of summary-worthy unigrams absent from the input (novel unigrams) was less than 15%. Finally, we removed references of figure from summaries and input, along with full tables from the input.

Salient Content Distribution (bigrams and longest common subsequences). As also shown in the main paper, i.e., Figure 4 and Figure 5, BIGPATENT demonstrates a relatively uniform distribution of the salient content from the summary

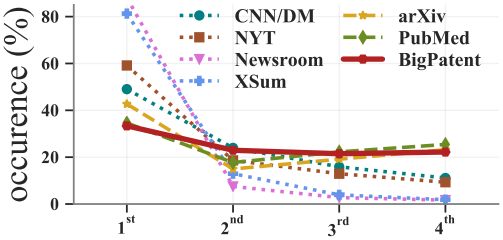

Figure 4: % of salient bigrams present in N^{th} segment of input.

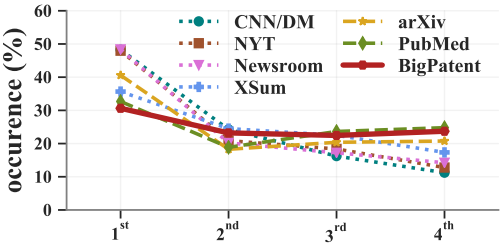

Figure 5: % of salient longest common subsequences present in N^{th} segment of input.

in all parts of the input. Here, the salient content is considered as all bigrams and longest common sub-sequences from the summary.

A.2 Experiment details

For all experiments, we randomly split BIGPATENT into 1,207,222 training pairs, 67,068 validation pairs, and 67,072 test pairs. For CNN/DM, we followed preprocessing steps from See et al. (2017), using 287,226 training, 13,368 validation, and 11,490 test pairs. For NYT, following preprocessing steps from Paulus et al. (2017), we used 589,298 training, 32,739 validation, and 32,739 test pairs.

Extract-based Systems. For TEXTRANK, we used the *summanlp*[7] (Barrios et al., 2016) to generate summary with three sentences based on TEXTRANK algorithm (Mihalcea and Tarau, 2004). For LEXRANK and SUMBASIC, we used *sumy*[8]. For RNN-EXT RL from Chen and Bansal (2018), we used the implementation provided by the authors[9].

Abstract-based Systems. For all the neural abstractive summarization models (except for SENTREWRITING), we truncated the input to 400 words and output to 100 words. Except for SENTREWRITING, all other models were trained us-

[7]https://pypi.org/project/summa/
[8]https://pypi.python.org/pypi/sumy
[9]https://github.com/ChenRocks/fast_abs_rl

ing *OpenNMT-py* python library[10] based on the instructions provided by the authors (Gehrmann et al., 2018b). We provide further details for each model below.

SEQ2SEQ with attention (Sutskever et al., 2014) was trained using a 128-dimensional word-embedding and 512-dimensional 1-layer LSTM. We used a bidirectional LSTM for the encoder and attention mechanism from Bahdanau et al. (2014). The model was trained using Adagrad (Duchi et al., 2011) with learning rate 0.15 and an initial accumulator value of 0.1. At inference time, we used the beam size 5. We used the same settings for training POINTGEN and POINTGEN + COV (See et al., 2017), adding the copy attention mechanism that allows the model to copy words from the source. At inference time, for POINTGEN + COV, we used coverage penalty with beta set to 5 and length penalty (Wu et al., 2016) with alpha as 0.9.

For SENTREWRITING from Chen and Bansal (2018), we again used the implementation by the authors[11] to train their full RL-based model using their default parameters.

A.3 Summaries for sample Input Document from BIGPATENT

For the sample summary presented in introduction of the main paper, in Table 7 we list complete gold-standard summary along with the summaries generated by SEQ2SEQ, POINTGEN + COV and SENTREWRITING. For the respective input, we also list the first 400 words for brevity.

[10] https://opennmt.net/OpenNMT-py/Summarization.html
[11] https://github.com/ChenRocks/fast_abs_rl

Gold-Standard summary
a shoelace cover incorporating an interchangeable fashion panel for covering the shoelaces of a gym shoe. the shoelace cover is secured to the shoe by a number of straps threaded through slots in the shoelace cover. a strap secured to each side of the gym shoe includes a loop and hook material such that the straps can be disengaged and the shoelace cover can be drawn back to expose the shoelaces of the shoe. the fashion panel is attached to the shoelace cover by a loop and hook material such that at the whim of the shoe wearer, the fashion panel can be replaced by other fashion panels to convey a fashion statement.
SEQ2SEQ generated summary
a shoe having a sole portion and an upper portion. the sole portion includes an upper portion and a lower portion. the upper portion is configured to receive the upper portion of the sole portion. the lower portion of the upper portion is configured to receive the upper portion of the sole portion.
POINTGEN + COV generated summary
a gym shoe and associated shoelace shoe is disclosed. the shoe includes a sole portion, a shoelace cover, and an upper portion. the upper portion has a toe area that extends from the toe area to the opening. the shoelace cover is curved to the shoelace.
SENTREWRITING generated summary
a gym shoe and associated shoelace cover and associated shoelace cover and fashion panel are disclosed. the shoe includes a sole portion and an upper portion. the shoelace cover is a semi-rigid panel that is curved to conform to the shoelace area of the shoelace area. the shoelace area is generally split into a shoelace area and a shoelace area. a shoe for use in a shoe, such as a shoe, is disclosed. a tongue extends from the toe area to the shoelace.
Input (first 400 words)
the following discussion of the preferred embodiment concerning a gym shoe and associated shoelace cover and fashion panel is merely exemplary in nature and is in no way intended to limit the invention or its applications or uses. the shoe includes a sole portion, generally comprised of a rugged rubber material, and an upper portion 14 generally comprised of a durable and pliable leather or canvas material. at a back location of the upper portion is an opening for accepting a wearer's foot. a cushion is visible through the opening on which the wearer's foot is supported. at a front end of the upper portion is a toe area. extending from the toe area to the opening is a shoelace area. the shoelace area is generally split such that a shoelace is threaded through eyelets associated with the shoelace area in order to bind together the shoelace area and secure the shoe to the wearer's foot. a tongue, also extending from the toe area to the opening, is positioned beneath the shoelace such that the tongue contacts the wearer's foot, and thus provides comfort against the shoelace to the wearer. the basic components and operation of a gym shoe is well understood to a person of normal sensibilities, and thus, a detailed discussion of the parts of the shoe and their specific operation need not be elaborated on here. secured to the upper portion of the shoe covering the shoelace area is a shoelace cover. in a preferred embodiment, the shoelace cover is a semi-rigid panel that is curved to be shaped to conform to the shoelace area such that an upper portion of the shoelace cover extends a certain distance along the sides of the upper portion adjacent the opening. the shoelace cover narrows slightly as it extends towards the toe area. the specifics concerning the shape, dimensions, material, rigidity, etc. of the shoelace cover will be discussed in greater detail below. additionally, the preferred method of securing the shoelace cover to the shoe will also be discussed below. in a preferred embodiment, affixed to a top surface of the shoelace cover is a fashion panel. the fashion panel is secured to the shoelace cover by an applicable securing mechanism, such as a loop and hook and/or velcro type fastener device, so that the fashion panel can be readily removed from the shoelace cover and replaced with an alternate fashion panel having a different design.

Table 7: Gold-standard and system generated summaries for BIGPATENT. Input (pre-processed) is truncated to 400 words for brevity.