WikiSum: Coherent Summarization Dataset for Efficient Human-Evaluation

Nachshon Cohen[*]
Amazon
nachshonc@gmail.com

Oren Kalinsky[*]
Amazon
orenk@amazon.com

Yftah Ziser[*]
Facebook[†]
yftahz@fb.com

Alessandro Moschitti
Amazon
amosch@amazon.com

Abstract

Recent works have made significant advances on summarization tasks, facilitated by summarization datasets. Several existing datasets have the form of coherent-paragraph summaries. However, these datasets were curated from academic documents written for experts, making the essential step of assessing the summarization output through human-evaluation very demanding.

To overcome these limitations, we present a dataset[1] based on article summaries appearing on the WikiHow website, composed of how-to articles and coherent-paragraph summaries written in plain language. We compare our dataset attributes to existing ones, including readability and world-knowledge, showing our dataset makes human evaluation significantly more manageable and effective. A human evaluation conducted on PubMed and the proposed dataset reinforces our findings.

1 Introduction

Summarization is the task of preserving the key information in a text while reducing its length. Recently, many summarization datasets were published and helped push the boundaries of new summarization systems. These datasets differ on several properties, including the domain (e.g., academic or news) and the summary form. PubMed, arXiv, and BigPatent (Cohan et al., 2018; Sharma et al., 2019) provide a summary in the form of coherent paragraphs (i.e., each sentence flows smoothly into the next). In contrast, other summarization datasets (Hermann et al., 2015; Grusky et al., 2018; Koupaee and Wang, 2018; Ladhak et al., 2020) offer a summary in the form of a key points list (i.e., highlights). In this paper, we focus on coherent paragraph summarization datasets.

How to Bake Chicken Breast? To bake chicken breast, start by lining a baking dish with foil or parchment paper. Then, put the chicken in the baking dish and bake it for 30-40 minutes at 400 degrees Fahrenheit, or until it reaches an internal temperature of 160 degrees Fahrenheit.
How to Break Up with Your Friend? The best way to break up with a friend is to confront them. Choose a time and place to meet up and explain to them why you are ending the friendship. Allow your friend to speak their mind as well, and work together to set boundaries for moving forward.

Figure 1: Examples of how-to questions and their corresponding answer's summarization in WikiSum.

Automatic evaluation of summarization systems, e.g., by using the ROUGE metric, is challenging (Lloret et al., 2018) and is often inconsistent with human evaluation (Liu and Liu, 2008; Cohan and Goharian, 2016; Tay et al., 2019; Huang et al., 2020). To understand – and later improve – the quality of summarization systems, it is necessary to conduct a human evaluation. A human evaluation's quality depends on the ease of reading and understanding of the measured text: a simple text does not require annotators with unique expertise, can be evaluated faster, and is easier to annotate correctly. However, existing coherent-paragraph summarization datasets consist of academic papers and cannot be considered easy to read. Evaluating such summarization samples requires unique expertise, takes time, and comes at a high cost.

In this work, we present WikiSum, a new summarization dataset from the WikiHow knowledge base[2]. The WikiSum documents are written in simple English, and the summaries provide "non-obvious tips that mimic the advice a knowledgeable, empathetic friend might give."[3] Unlike previous WikiHow summarization (Koupaee and Wang, 2018; Ladhak et al., 2020) datasets and summaries

[*] Co-first author

[†] Work done while at Amazon

[1] The dataset and human evaluation are available at https://registry.opendata.aws/wikisum.

[2] https://www.wikihow.com

[3] https://www.wikihow.com/Write-or-Edit-a-Quick-Summary-on-wikiHow

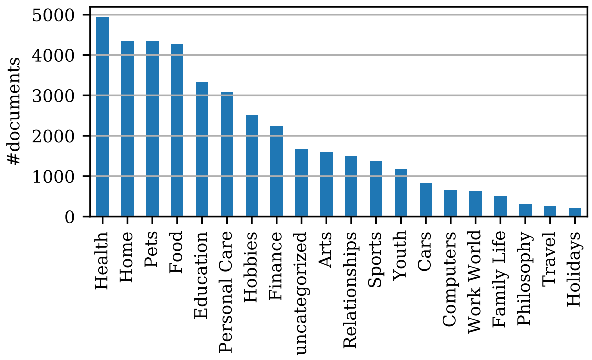

Figure 2: Category distribution in WikiSum.

from the news domain, the summaries of WikiSum are in the form of a coherent paragraph written by the document authors (examples in Figure 1). Moreover, in contrast to other coherent-paragraph summarization datasets from the academic domain, WikiSum is written using simple English. This critical property can help with the challenging task of evaluating summarization systems and provide insights that can go unnoticed using automatic evaluation methods.

The key attributes of WikiSum are: (1) Summaries written as a single, coherent passage. (2) Articles and summaries that are easy to read. (3) Articles and summaries require less world knowledge to understand. We evaluate the dataset readability and estimate the required world-knowledge in Section 3. Moreover, we reinforce our results by conducting a human-evaluation of a summarization dataset in Section 4. Finally, to establish a baseline on the proposed dataset, we benchmark WikiSum using recent summarization systems and report their performance on Section 5.

2 Related Work

The summarization landscape can be roughly divided into three primary summary-forms: (1) Single sentence (Napoles et al., 2012; Grusky et al., 2018; Narayan et al., 2018; Kim et al., 2019) - summarize the document in a single sentence; (2) Highlights (Hermann et al., 2015; Koupaee and Wang, 2018; Ladhak et al., 2020) - a summary in the form of bullets listing the key points in the text; (3) Coherent summary (Sharma et al., 2019; Cohan et al., 2018) - short coherent paragraphs describing the salient information. The summarization datasets from the news domain, which are commonly used for human evaluation, include summaries in the form of highlights or single-sentence summaries. However, summarization datasets written in a coherent format come from the academic domain, making them extremely difficult to annotate manually. Our proposed WikiSum is the only dataset written in a coherent format, yet easy for human evaluation. We do not claim that coherent paragraph summaries are *better*, but rather *different*; each format has its use cases, and human evaluation should be done on each of the different formats separately.

The existing WikiHow datasets (Koupaee and Wang, 2018; Ladhak et al., 2020) can be considered the closest to WikiSum, as they originate from the same knowledge base. However, while the existing WikiHow datasets split the article to generate the document and summary, WikiSum uses the entire article as the document and a summary specifically written by the article's author (called the Article Quick Summary). The former uses the concatenation of the first line of each step, called the step header, as the list of highlights and the remainder of step text's concatenation called ”wrap-text,” as the document[4]. In addition to the different summary-form of the highlight-based WikiHow and WikiSum, the content of the summaries is significantly different, which can be illustrated by the low BLEU-4 (0.06[5]) between the two.

BigPatent (Sharma et al., 2019), Arxiv and PubMed (Cohan et al., 2018) are recent summarization datasets with coherent paragraph summaries. These datasets focus on the academic domain and are written for experts. Like these datasets, WikiSum is composed of long documents and coherent paragraph summaries. Nonetheless, it uses common everyday language and ranges over many domains (see Figure 2). Finally, Table 1 compares WikiSum to common existing datasets. Additional details on WikiSum are available in the appendix.

3 Measuring Text Difficulty

This section focuses on two crucial attributes: ease of readability and external knowledge required, shown (in Section 4) to be important for easy and effective human evaluation. For brevity, we focus on summarization datasets with coherent-paragraph summaries.

[4] WikiHow author instructions (wikihow.com, 2020) specifically states that the authors can use the wrap-text to describe why the step header is important. This leads to many cases where the step headers are not a summary of the wrap-text.

[5] We used WikiSum as the reference, the results are very similar when WikiHow is used as a reference. ROUGE-1, 2 and L are 0.37, 0.13, and 0.23, respectively.

	Domain	# Docs	Comp. ratio	Summary # word	Summary # sent	Doc # word
WikiSum	instructional	39,775	13.9	101.2	5.0	1,334.2
arXiv	academic	215,913	39.8	292.8	9.6	6,913.8
PubMed	academic	133,215	16.2	214.4	6.9	3,224.4
BigPatent	academic	1,341,362	36.4	116.5	3.5	3,572.8
WikiHow	instructional	215,365	14.5	69.0	7.2	500.8
CNN/DM	news	312,085	13.0	55.6	3.8	789.9
NYT	news	654,788	12.0	44.9	2.0	795.9
Newsroom	news	1,212,726	43.0	30.4	1.4	750.9
XSum	news	226,711	18.8	23.3	1.0	431.1

Table 1: Statistics comparison of summarization datasets. Datasets not in coherent-paragraph form are marked in gray.

3.1 Readability

Readability metrics attempt to indicate how difficult a passage in English is to read. We used classical readability measures, including FKGL (Farr et al., 1951), GFI (Robert, 1968), SMOG (Mc Laughlin, 1969), ARI (Senter and Smith, 1967), CLI (Coleman and Liau, 1975). All these metrics are based on lexical features of the text, e.g., number of words in a sentence or mean number of syllables per word. They produce a score that is interpreted as the number of years of formal education required (for a native English speaker) to understand a piece of text[6].

For each document, we measured readability scores[7] for the document and the ground truth summary. The document is longer than the summary, so its readability is of higher importance. We report the average readability score for all the samples in the dataset.

Readability scores for the documents are presented at the top of Table 2. The table shows that WikiSum is significantly easier to read than other documents from coherent-summary datasets (arXiv, PubMed, BigPatent). Similar results can be found for the readability scores for the summaries (bottom of Table 2). To conclude, WikiSum is measured as drastically simpler to read than other coherent-summary datasets.

3.2 External Knowledge

Existing datasets are composed of academic documents that are written for experts. Often, to fully understand academic texts requires domain knowledge, which makes the annotator pool smaller, and thus, in most cases, more expensive. Word frequency is a strong indicator of how familiar a word is (Paetzold and Specia, 2016), where rare words tend to be less familiar.

	Dataset	ARI	FKGL	GFI	SMOG	CLI
Document	WikiSum	7.4	6.82	10.15	9.71	8.83
	arXiv	14.02	13.51	18.47	15.44	14.31
	PubMed	16.74	16.27	20.64	17.03	15.01
	BigPatent	13.46	13.32	17.47	14.68	11.68
Summary	WikiSum	9.71	8.49	11.91	10.24	8.78
	arXiv	16.44	16.1	20.5	16.8	15.23
	PubMed	17.73	17.35	21.6	17.44	16.6
	BigPatent	22.47	20.91	25.12	18.75	14.0

Table 2: Readability scores for the documents (top) and summaries (bottom), measured in years of formal education required to read the text. Smaller is simpler.

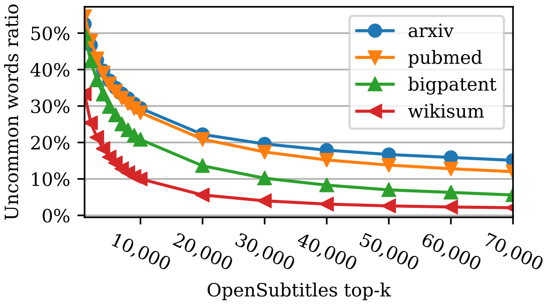

Figure 3: Ratio of uncommon words in the document, which cannot be found in the Top-K OpenSubtitles words, for different k values.

We used OpenSubtitles (Lison and Tiedemann, 2016), text corpora compiled from an extensive database of movie and TV subtitles to obtain word frequencies. We hypothesize that movie and TV subtitles can roughly represent common knowledge among many people. In Figure 3, we show the percentage of non-frequent words in a document (i.e., words that cannot be found in the top-k words in OpenSubtitles) as a function K, averaged over a random sample of $10,000$ documents from each dataset. This figure clearly shows that WikiSum is composed of significantly fewer words unpopular in TV shows and movies, requiring less specialized external knowledge.

4 Human Evaluation

We conducted a standard human evaluation on a summarization task, in addition to the automatic readability and the external knowledge metrics. We gathered a pool of 6 annotators, without any prior knowledge of the project, all with a graduate degree (M.sc. or Ph.D.) and proficient English reading-level. We asked them to evaluate summaries generated by Pegasus (Zhang et al., 2020). The an-

[6] Other readability metrics such as FRE (Flesch, 1948), LIX and RIX (Björnsson, 1968), have a similar trend to the shown metrics, but require a translation to years of education, omitted from this paper for brevity.

[7] https://github.com/mmautner/readability

dataset	time (minutes)	difficulty (rating)	exhausting (rating)	qualified (rating)	unknown (%)
WikiSum	6.8±1.2	1.9±0.3	2.2±0.5	4.2±0.3	0.2±0.1
PubMed	10.0±1.2	3.7±0.3	3.9±0.4	2.2±0.4	3.7±1.4

Table 3: Evaluation time per sample, evaluation difficulty/exhaustion rating, perceived qualification, and the ratio of unknown words in the document. ± denotes 95% confidence interval according to student's t distribution (df=20). Difficulty, qualification, and tiring were marked on a 1-5 scale.

Models	LEAD-3	TextRank	PEGASUS$_{\text{LARGE}}$
WikiSum	25.3/6.84/16.2	32.7/8.8/18.9	43.35/15.48/26.91
arXiv	25.53/5.98/15.22	33.1/9.7/18.1	43.07/19.70/34.79
PubMed	26.38/8.73/16.6	35.3/13.1/20.4	44.70/17.27/25.80
BigPatent	28.9/7.96/18.17	33.0/9.8/19.6	45.49/19.90/27.69

Table 4: ROUGE-1/2/L F1 scores on coherent-summary datasets. Pegasus baseline results are from (Zhang et al., 2020), except for WikiSum.

notation task followed Huang et al. (2020) and consisted of relevance, consistency, fluency, and coherency.

Due to resource limitations (and the difficulty of annotating articles from the academic domain), we had to pick one coherent-paragraph dataset for comparison with WikiSum. To avoid annotators' domain bias, we selected articles from PubMed, which contains articles not in the area of expertise of any annotator, in addition to WikiSum. We sampled random articles with 950 - 1050 words to avoid length bias, ensuring that article length is similar in both datasets. All annotators allocated 1 hour, which amounted to 42 annotations, 21 for each dataset.

During the annotation task, we measured the evaluation time and asked the annotators to mark unfamiliar words. In addition, we asked the annotators to rate the following aspects on a 1-5 scale: (a) How difficult was the task? (b) How tiring was it? (c) How qualified are you for this task? After each pair of PubMed and WikiHow samples were completed, the annotators selected which dataset they prefer to evaluate.

In Table 3 we show the annotators' assessment of the tasks. Compared to PubMed, a WikiSum annotation takes significantly less time, is less difficult, and less tiring. Moreover, the annotators revealed that they were much more qualified to assess the WikiSum task summary. Finally, in 90% of the cases (19 out of 21), the annotators revealed that they preferred a WikiSum annotation task. This reinforces our findings that WikiSum is significantly easier to annotate than PubMed.

In the annotation task, we also asked the annotators to mark unfamiliar words in the article. We found a strong correlation between the count of unfamiliar words and the task difficulty, evaluation time, and perceived required qualification (Pearson correlation of $0.57, 0.36, -0.48$[8], respectively, $p < 0.05$). Strong correlation was also found between the ARI readability metric (Section 3.1) and the above-mentioned annotation metrics (Pearson correlation of $0.69, 0.49, -0.76$, $p < 0.05$). This demonstrates the effect of readability on the difficulty of an annotation task.

Finally, we found that unfamiliar words correspond to low-frequency OpenSubtitles words (Section 3.2). The *unfamiliar* words on WikiSum and PubMed appear in the top $91,550$ and $230,596$ words on average, respectively, while *familiar* words appear in the top $16,935$ and $59,244$ words on average, respectively. It also further validates Paetzold and Specia (2016) hypothesis about the strong correlation between word frequency and complexity.

5 Model Results and Discussion

To provide both abstractive and extractive baselines for WikiSum, we evaluate on PEGASUS$_{\text{LARGE}}$ (Zhang et al., 2020), TextRank (Mihalcea and Tarau, 2004), and the common LEAD-3 that selects the first three sentences of the document as the summary. We compare the results on WikiSum to the Arxiv, PubMed, and BigPatent Datasets results. Table 4 reports the F1 scores of ROUGE-1, 2 and L for all the models. The results show that the models' performance on WikiSum is not drastically different from the other datasets, making it an interesting dataset for benchmarking summarization systems. The detailed evaluation setup can be found in the supplementary materials.

To conclude, this paper presents the WikiSum dataset, which is drastically simpler for human evaluation than existing summarization datasets where the summary appears as a coherent paragraph. We showed WikiSum's simplicity via various readability metrics and demonstrated that the text requires less external knowledge to be understood. Finally, we validated our finding via a human evaluation task on WikiSum and PubMed.

[8] Many unfamiliar words implied annotators perceived themselves as less unqualified.

References

Carl Hugo Björnsson. 1968. Läsbarhet, stockholm: Liber.

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, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 615–621. Association for Computational Linguistics.

Arman Cohan and Nazli Goharian. 2016. Revisiting summarization evaluation for scientific articles. *CoRR*, abs/1604.00400.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

James N Farr, James J Jenkins, and Donald G Paterson. 1951. Simplification of flesch reading ease formula. *Journal of applied psychology*, 35(5):333.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

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, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 708–719. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kociský, 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: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469, Online. Association for Computational Linguistics.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2019. Abstractive summarization of reddit posts with multi-level memory networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2519–2531. Association for Computational Linguistics.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *CoRR*, abs/1810.09305.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen R. McKeown. 2020. Wikilingua: A new benchmark dataset for multilingual abstractive summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, pages 4034–4048. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Feifan Liu and Yang Liu. 2008. Correlation between ROUGE and human evaluation of extractive meeting summaries. In *Proceedings of ACL-08: HLT, Short Papers*, pages 201–204, Columbus, Ohio. Association for Computational Linguistics.

Elena Lloret, Laura Plaza, and Ahmet Aker. 2018. The challenging task of summary evaluation: an overview. *Language Resources and Evaluation*, 52(1):101–148.

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*, pages 404–411. ACL.

Courtney Napoles, Matthew R. 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@NAACL-HLT 2012, Montrèal, Canada, June 7-8, 2012*, 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, Brussels, Belgium, October 31 - November 4, 2018*, pages 1797–1807. Association for Computational Linguistics.

Gustavo Paetzold and Lucia Specia. 2016. Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.

Gunning Robert. 1968. The technique of clear writing. *Revised Edition). New York: McGraw Hill.*

RJ Senter and Edgar A Smith. 1967. Automated readability index. Technical report, CINCINNATI UNIV OH.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2204–2213. Association for Computational Linguistics.

Wenyi Tay, Aditya Joshi, Xiuzhen Jenny Zhang, Sarvnaz Karimi, and Stephen Wan. 2019. Red-faced rouge: Examining the suitability of rouge for opinion summary evaluation. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 52–60.

wikihow.com. 2020. WikiHow Article Guidelines. https://www.wikihow.com/Write-a-New-Article-on-wikiHow.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

A Data Description

A.1 Gathering the data

We use Scrapy scraper[9] to download articles and summaries from the wikihow.com website. We removed HTML tags using BeautifulSoup[10]. Finally, we removed any sample in which the summary is a list of bullet points; around 7k samples were excluded in this manner.

A.2 Authors Instructions for Writing Quick Summaries

The wikihow.com website provides the following guidelines for authors writing a quick summary.[11]

> The goal of the “Quick Summary” section on wikiHow is to provide a short summary of non-obvious tips that mimic the advice a knowledgeable, empathetic friend might give you if you asked them for help on the given topic. Among other uses, Quick Summaries help smart devices like Google Homes and Amazon Echos deliver wikiHow advice to listeners in need of how-to guidance.

We remark that the quick summaries are indeed used by commercial voice assistants to answer how-to questions. As voice assistants gain popularity, so does the importance of such coherent-paragraph summaries.

A.3 Data Layout

Raw data is available in the supplementary material, in a json format. Each line consists of a single sample, with the following fields

1. Link to the original article
2. Article title
3. Article text
4. Quick summary
5. Split fold (train, dev, or test)

Finally, it also includes *step_headers*: the first line in each step. This is part of the article but might be considered more important, and therefore, it might find further uses by system designers.

A.4 Dataset Statistics

Most dataset statistics appear in Table 1 in the article's main body and are repeated here for completeness. The total number of samples in the WikiSum dataset is $39,775$. On average, each summary consists of 101.2 words, while each article consists of $1,334.2$ words. The average compression ratio is 13.9.

A.5 Evaluation details

We randomly split WikiSum into 35,775 (document, summary) training pairs, as well as 2,000 validation pairs and 2,000 test pairs. The rest of the datasets were downloaded from the HuggingFace dataset repository[12].

All the datasets were evaluated using TextRank[13] and Pegasus-large. The ROUGE scores throughout the paper were calculated using rouge-score[14]. We utilized TextRank to generate three summary sentences. The Pegasus results on Arxiv, Pubmed, and Arxiv were taken from the Pegasus paper. The results on WikiSum were computed by using the Github repository of the Pegasus paper[15]. Pegasus was trained on a single NVIDIA V100 Tensor Core

[9] www.scrapy.org

[10] https://pypi.org/project/beautifulsoup4

[11] https://www.wikihow.com/Write-or-Edit-a-Quick-Summary-on-wikiHow

[12] https://huggingface.co/datasets

[13] https://pypi.org/project/summa

[14] https://pypi.org/project/rouge-score/

[15] https://github.com/google-research/pegasus

GPU, using max input and output sequence lengths of 1024 and 256, respectively.

B Example Summaries

In this appendix, we provide an example summary from WikiSum and arXiv, PubMed, and bigPatent. Note that the article can be quite long (for arXiv and PubMed, it is a full academic paper), so it is not presented in this appendix. Instead, we provide a link to the online version of the full article.

B.1 WikiSum

The WikiSum example summary is provided below:

"To ace a test, even if you're not prepared, start by glancing over the test before you get started to get an idea of how long it is so you can manage your time better. Then, read through each question twice and try to answer it. If you can't answer a question, skip it and come back to it later if you can, which will save you from wasting all of your time on one question. If your test is multiple choice and you don't know the answer, eliminate two answers, so you're left with just two options. Then, guess if necessary since you'll have a 50-percent chance of being right."

The article is available at https://www.wikihow.com/Ace-a-Test.

B.2 WikiHow

For the sake of comparison between WikiHow and WikiSum datasets, we provide the WikiHow summary originating from the same raw material (i.e., the same wikihow.com how-to article) as the WikiSum example at Appendix B.1. We remark that the article to be summarized is not exactly the same, as the WikiHow example does not contain the step headers from the article's text. The WikiHow summary is provided below.

"Study well before the test. Get a study friend. Take breaks. Relax. Pay attention in class. Do all available practice questions. Get some sleep the night before. Have proper meals before the test day. Have your test-taking materials assembled and ready. Listen to music you like. Go into the test in a positive manner. Take deep breaths to try to keep calm. Read the questions carefully. Do the easy questions first. Go with your first answer. Use logic if you're stuck on a multiple choice question. Review your answers thoroughly when you are done."

It can easily be seen that the WikiSum summary is a coherent, fluent paragraph, while the WikiHow summary is a set of bullet points. The content of the two summaries are also quite different between the two datasets.

B.3 arXiv

"the effect of a random phase diffuser on fluctuations of laser light (scintillations) is studied. not only spatial but also temporal phase variations introduced by the phase diffuser are analyzed. the explicit dependence of the scintillation index on finite - time phase variations is obtained for long propagation paths. it is shown that for large amplitudes of phase fluctuations , a finite - time effect decreases the ability of phase diffuser to suppress the scintillations."

The article is available at https://arxiv.org/pdf/0903.5449.pdf.

B.4 PubMed

"tardive dystonia (td) is a serious side effect of antipsychotic medications, more with typical antipsychotics, that is potentially irreversible in affected patients. studies show that newer atypical antipsychotics have a lower risk of td. as a result, many clinicians may have developed a false sense of security when prescribing these medications. we report a case of 20-year - old male with hyperthymic temperament and borderline intellectual functioning, who developed severe td after low dose short duration exposure to atypical antipsychotic risperidone and then olanzapine. the goal of this paper is to alert the reader to be judicious and cautious before using casual low dose second generation antipsychotics in patient with no core psychotic features, hyperthymic temperament, or borderline intellectual functioning suggestive of organic brain damage, who are more prone to develop adverse effects such as td and monitor the onset of td in patients taking atypical antipsychotics."

The article is available at
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5330001/.

B.5 BigPatent

"this invention relates to novel calcium phosphate - coated implantable medical devices and processes of making same. the calcium - phosphate coatings are designed to minimize the immune response to the implant and can be used to store and release a medicinally active agent in a controlled manner.

such coatings can be applied to any implantable medical devices and are useful for a number of medical procedures including balloon angioplasty in cardiovascular stenting, ureteral stenting and catheterisation. the calcium phosphate coatings can be applied to a substrate as one or more coatings by a sol - gel deposition process, an aerosol - gel deposition process, a biomimetic deposition process, a calcium phosphate cement deposition process, an electro - phoretic deposition process or an electrochemical deposition process. the coating can contain and elude a drug in an engineered manner."

The article is available at https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2007147234.