www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

Question-driven summarization of answers to consumer health questions

Max Savery, Asma Ben Abacha, Soumya Gayen & Dina Demner-Fushman

Automatic summarization of natural language is a widely studied area in computer science, one that is broadly applicable to anyone who needs to understand large quantities of information. In the medical domain, automatic summarization has the potential to make health information more accessible to people without medical expertise. However, to evaluate the quality of summaries generated by summarization algorithms, researchers first require gold standard, human generated summaries. Unfortunately there is no available data for the purpose of assessing summaries that help consumers of health information answer their questions. To address this issue, we present the MEDIQA-Answer Summarization dataset, the first dataset designed for question-driven, consumer-focused summarization. It contains 156 health questions asked by consumers, answers to these questions, and manually generated summaries of these answers. The dataset's unique structure allows it to be used for at least eight different types of summarization evaluations. We also benchmark the performance of baseline and state-of-the-art deep learning approaches on the dataset, demonstrating how it can be used to evaluate automatically generated summaries.

Background & Summary

A summary is a concise description that captures the salient details from a more complex source of information[1]. Summaries are regularly used as a tool to quickly understand content from a single source, such as a book or movie, or from many disparate sources, such as news stories about a recent event. Even this article began with a summary: an abstract.

Summarization can be particularly useful for helping people easily understand online heath information. One of the first places people turn to for answers to their health questions is the internet[2]. A conventional search engine will return a set of web pages in response to a user's query, but without considerable medical knowledge the consumer is not always able to judge the correctness and relevance of the content[3]. In fact, finding relevant biomedical material can be difficult for even medical experts[1]. While having a reliable, easy-to-understand summary for an article—such as one similar to the plain language summaries created by the health organization Cochrane[4]—would likely make searching for health information easier, it is not possible to tailor a manually generated summary to every user. For this reason, a summary automatically generated in response to a user's query could be extremely beneficial, especially for users who do not have medical expertise.

Recent developments in automatic text summarization, a field at the intersection of machine learning and natural language processing (NLP), have shown the potential to aid consumers in understanding health information[5]. However, to develop more advanced summarization algorithms capable of reliably summarizing medical text, researchers require human curated datasets that can be used to consistently measure the quality of machine generated summaries. Unfortunately, there is currently a lack of question-driven and consumer-focused data available, i.e., human generated summaries of information relevant to helping consumers answer their health questions. A dataset for this purpose must contain the following data: (1) questions asked by people without medical expertise; (2) documents containing answers to the questions; and (3) easily understood summaries that are informed by the health questions asked by consumers. In order to address the absence of data that meets these conditions, the contribution of this paper is as follows: A new gold standard dataset, MEDIQA-Answer Summarization (MEDIQA-AnS)[6], consisting of 156 health questions asked by consumers, corresponding answers to these questions, and expert-created summaries of these answers.

Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. ✉e-mail: ddemner@mail.nih.gov

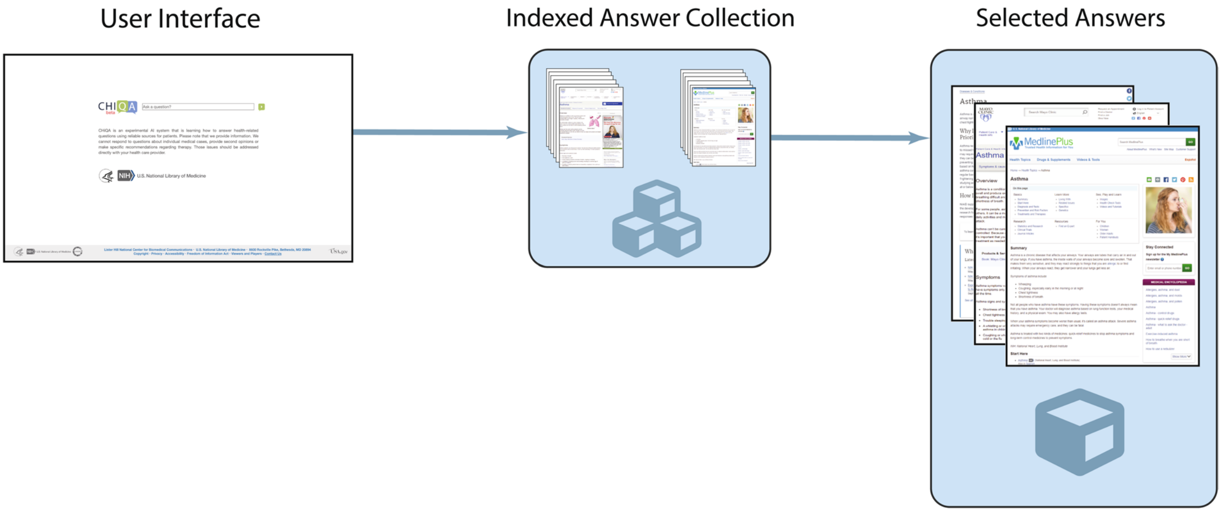


Fig. 1 CHiQA's user interface and answer processing pipeline.

There are many available summarization datasets, but none satisfy the conditions mentioned above. For example, popular summarization datasets include the CNN-Dailymail dataset[7], which uses headlines as summaries of news articles, and the PubMed dataset[8], which uses abstracts as summaries of scientific articles. These can be used for training and testing summarization systems on short, open domain text and long, scientific text. Additionally, Multi-News[9] can be used for multi-document summarization of news articles, and BioASQ[10] can be used for question-driven, single or multi-document summarization of biomedical text. Though the BioASQ data approaches the requirements for consumer health summarization, the questions and summaries are technical in nature.

Recently, the MEDIQA 2019 shared task[11] introduced the MEDIQA-QA dataset for answer-ranking, encouraging medical question answering research. MEDIQA-QA is uniquely suited for the purpose of this paper: It consists of consumer health questions and passages that contain information relevant to the question. This fulfills two of our three conditions for question-driven answer summarization. We therefore used the passages in MEDIQA-QA as the primary data source for MEDIQA-AnS. To extend MEDIQA-QA for summarization, we manually generated single and multi-document summaries of the passages. We also created two versions of each summary: An extractive version, consisting of content copied-and-pasted from the passages, and an abstractive version, written from scratch, using the passages as reference. This makes at least eight different types of summarization evaluations possible, including single document or multi-document summarization, on either long or short documents, and with extractive or abstractive approaches. Researchers will be able to evaluate models in a wide variety of summarization environments, whereas many previously published datasets can only be used for one or two types of evaluations. In addition to releasing MEDIQA-AnS, we include experiments using baseline and state-of-the-art summarization approaches, focusing on the single document aspect of the task, in order to benchmark the dataset for future researchers.

Methods

Data creation. The MEDIQA-AnS dataset introduced in this paper contains consumer health questions, the full text from reliable web pages, extracted passages from the full text, and manually created summaries. The questions in MEDIQA-AnS are a subset of those in MEDIQA-QA, consisting of questions submitted to the National Library of Medicine's Consumer Health information Question Answering (CHiQA)[2] online system, shown in Fig. 1. CHiQA indexes only pages hosted by reliable organizations, such as MedlinePlus and the Mayo Clinic. In response to consumers' health questions, it provides passages from these pages, using information retrieval and machine learning techniques. The MEDIQA-QA dataset uses these passages as answers to the associated questions. It also contains manual ratings of the relevance of the passages to the question.

To create the summaries in MEDIQA-AnS, we first filtered passages from MEDIQA-QA which had been rated as "relevant but incomplete" or "excellent". Table 1 shows the frequency of websites that this subset of answers was selected from. Then, for each question and corresponding set of answers, we generated the following types of summaries:

- Extractive summary of each answer
- Abstractive summary of each answer
- Multi-document extractive summary considering the information presented in all of the answers
- Multi-document abstractive summary

The summaries of the answers were generated by two subject matter experts, using a summarization interface we created to allow the annotators to input separate summaries of each type. The extractive summaries were generated by selecting chunks of text from the answers. Though the source text was sometimes punctuated correctly,

Website	Frequency
medlineplus.gov	190
mayoclinic.org	151
nlm.nih.gov	44
rarediseases.info.nih.gov	39
ghr.nlm.nih.gov	31
nhlbi.nih.gov	22
niddk.nih.gov	21
ninds.nih.gov	16
womenshealth.gov	12
nihseniorhealth.gov	8
nichd.nih.gov	7
niams.nih.gov	6
cancer.gov	3
nia.nih.gov	1
nei.nih.gov	1

Table 1. Frequency of reliable websites included in MEDIQA-AnS.

Data type	Text
Question	*What is the consensus of medical doctors as to whether asthma can be cured?*
Document	Asthma is a condition in which your airways narrow and swell and produce extra mucus. This can make breathing difficult and trigger coughing, wheezing and shortness of breath. […]
Summary	Asthma can't be cured, but its symptoms can be controlled. Because asthma often changes over time, it's important that you work with your doctor to track your signs and symptoms and adjust treatment as needed […]
Question	*hi, I would like to know if there is any support for those suffering with abetalipoproteinemia? I am not diagnosed but have had many test that indicate I am suffering with this, keen to learn how to get it diagnosed and how to manage, many thanks*
Document	Bassen-Kornzweig syndrome is a rare disease passed down through families. The person is unable to fully absorb dietary fats through the intestines. Causes Bassen-Kornzweig syndrome is caused by a defect in a gene […]
Summary	Abetalipoproteimemia, also known as Bassen-Kornzweig syndrome, is diagnosed using blood tests for Apolipoprotein B, vitamin deficiencies, malformation of red blood cels, complete blood count and cholesterol. […]

Table 2. Examples of questions, documents, and summaries in MEDIQA-Ans.

it also included lists, headings, and other types of formatting. We selected text regardless of formatting, considering which chunks contained the relevant information.

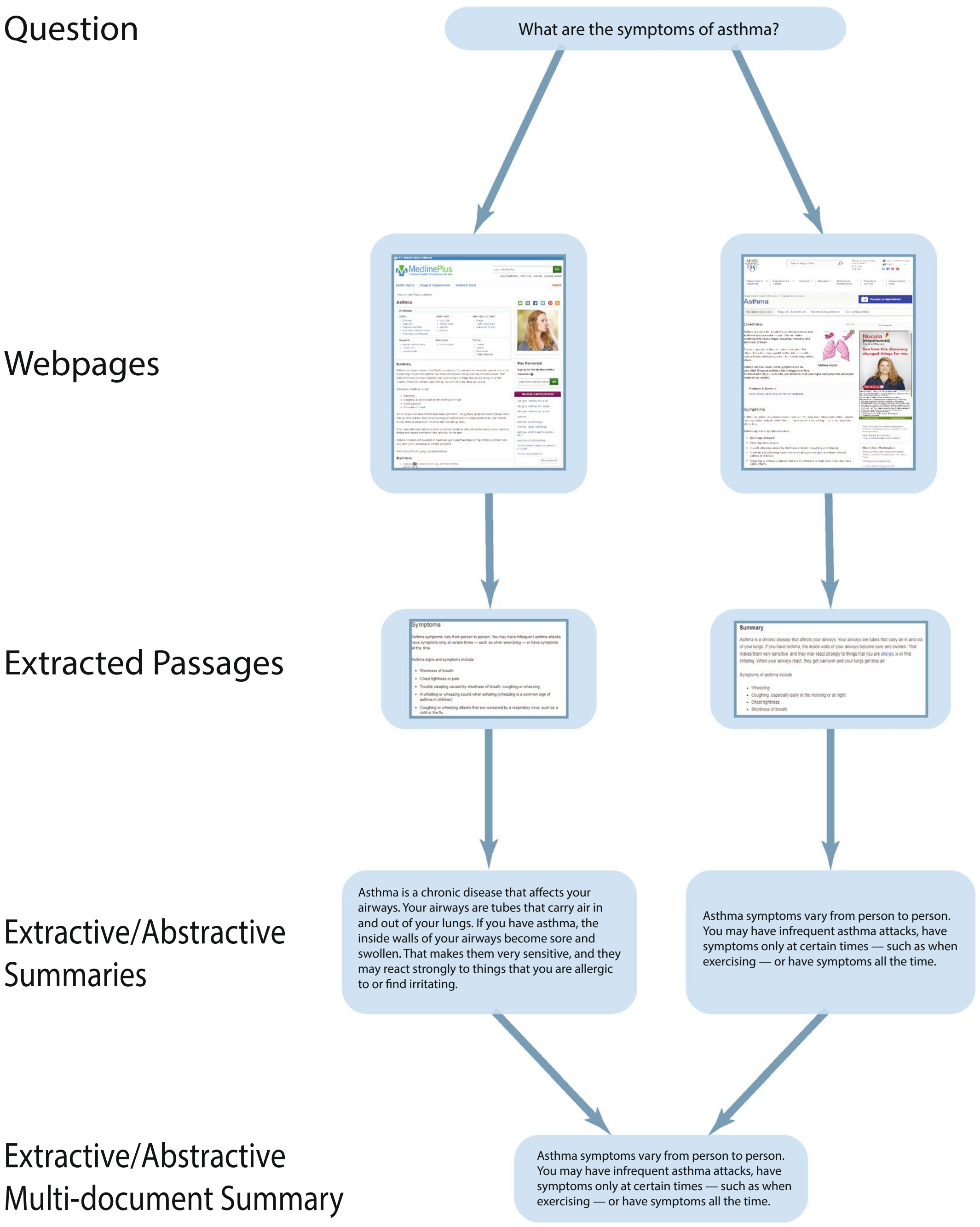
For the abstractive summaries, the source text was rewritten into easier to understand, condensed sentences. Writing the abstractive summaries involved either rewording chunks, reorganizing and reformatting sentences, or potentially using extracted text that was already clear and informative. Since the answers were selected from reliable online medical resources, there were many cases in which the extractive summary was already well-worded and clear. Finally, once the extractive and abstractive summaries were written, multi-document summaries were created using all of the answers. Examples taken from the dataset can be seen in the Table 2.

Evaluation metrics.

We use ROUGE[12] and BLEU[13], both widely-used measures of text similarity, to calculate agreement between the annotators, compare the extractive and abstractive summaries, and evaluate the automatically generated summaries. ROUGE and BLEU measure the number of contiguous words (referred to as n-grams in NLP) occurring in a candidate summary when compared to a reference summary. For example, ROUGE-2 measures the number of contiguous two word (bigram) sequences that occur in both the candidate and reference summary, penalizing the candidate for missing bigrams. This means that ROUGE is oriented for recall, and, conversely, BLEU is oriented for precision, penalizing the candidate for including incorrect n-grams. We report ROUGE-1, ROUGE-2, ROUGE-L, and BLEU-4. Note that ROUGE-L is computed slightly differently than the other ROUGE variations, measuring the longest common subsequence between a candidate and reference.

Data Records

We have archived nine data records with Open Science Framework (OSF), available at https://doi.org/10.17605/OSF.IO/FYG46[6]. The primary dataset contains 156 questions, the text of the web pages returned by CHiQA, the passages selected from the web pages, and the abstractive and extractive, multi and single document summaries of the passages. The additional eight datasets are subsets of the primary one, divided into potential experimental use cases. For example, we have included a split containing questions, corresponding web pages, and the multi-document summaries of these pages. This allows users to directly evaluate a system on multi-document summarization without having to perform additional data processing on the whole dataset. There are potentially more than eight use cases, if users are interested in using the passages as the summaries of the full text as a kind of long-form summary. However, we have not provided any pre-made splits of the data for this purpose.


Question
Webpages
Extracted Passages
Extractive/Abstractive Summaries
Extractive/Abstractive Multi-document Summary

What are the symptoms of asthma?

Symptoms

Asthma symptoms vary from person to person. You may have infrequent asthma attacks, have symptoms only at certain times — such as when exercising — or have symptoms all the time.

Asthma signs and symptoms include:

- Shortness of breath
- Chest tightness or pain
- Trouble sleeping caused by shortness of breath, coughing or wheezing
- A whistling or wheezing sound when exhaling (wheezing is a common sign of asthma in children)
- Coughing or wheezing attacks that are worsened by a respiratory virus, such as a cold or the flu

Summary

Asthma is a chronic disease that affects your airways. Your airways are tubes that carry air in and out of your lungs. If you have asthma, the inside walls of your airways become sore and swollen. That makes them very sensitive, and they may react strongly to things that you are allergic to or find irritating. When your airways react, they get narrower and your lungs get less air.

Symptoms of asthma include

- Wheezing
- Coughing, especially early in the morning or at night
- Chest tightness
- Shortness of breath

Asthma is a chronic disease that affects your airways. Your airways are tubes that carry air in and out of your lungs. If you have asthma, the inside walls of your airways become sore and swollen. That makes them very sensitive, and they may react strongly to things that you are allergic to or find irritating.

Asthma symptoms vary from person to person. You may have infrequent asthma attacks, have symptoms only at certain times — such as when exercising — or have symptoms all the time.

Asthma symptoms vary from person to person. You may have infrequent asthma attacks, have symptoms only at certain times — such as when exercising — or have symptoms all the time.

Fig. 2 Example of a single example in the primary MEDIQA-AnS data record.

Each dataset is saved in JSON format, where each key is a question id and, depending on the dataset, each value contains a nested JSON object with the question, text of the web pages, passages, summaries, rating of passages from MEDIQA-QA, and the URL for each web page. A summary of the structure of the data for a single example is shown in Fig. 2, and statistics regarding the questions, articles, and summaries are shown in Table 3. More detailed descriptions regarding the potential use of each dataset and their respective key and value pairs can be found in the README file in the OSF archive.

Technical Validation

Inter-annotator agreement. For a subset of questions in MEDIQA-AnS, the annotators summarized the same passages, so that inter-annotator agreement between the respective summaries could be calculated. The ROUGE and BLEU scores measuring the similarity are shown in Table 4. It is apparent that the annotators more frequently use the same n-grams when creating extractive summaries. This is to be expected, as it is less likely that two individuals will use the exact same combinations of words when generating novel, abstractive text.

Data type	Count	Words		Sentences	
		Average	S.d.	Average	S.d.
Questions	156	25	31	2	2
Unique articles	348	1675	1798	95	104
Passages	552	631	869	35	48
Summaries					
Multi-document abstractive	156	141	119	7	6
Multi-document extractive	156	220	183	12	12
Single-document abstractive	552	83	78	4	4
Single-document extractive	552	133	127	7	8

Table 3. Average number and standard deviation of words and sentences per document type.

Summarization type	ROUGE-2	ROUGE-L	BLEU
Multi-document, abstractive	0.19	0.32	0.17
Multi-document, extractive	0.56	0.57	0.49
Single document, abstractive	0.28	0.42	0.19
Single document Extractive	0.82	0.83	0.74

Table 4. ROUGE-2, ROUGE-L and BLEU inter-annotator agreement for each summary type.

Summary type	ROUGE-2	BLEU
Single document, abstractive v. extractive	0.64	0.41
Multi-document abstractive v. extractive	0.62	0.42

Table 5. ROUGE-2 and BLEU calculated using abstractive summaries as the reference summary and the extractive summaries as the candidate summary.

Additionally, we wanted to measure the similarity between the abstractive and extractive summaries. Using the abstractive summaries as the reference summary and the extracted summaries as the candidates, the scores shown in Table 5 indicate that the extractive summaries do contain many of the same n-grams as the abstractive summaries.

However, it is important to note that even if a pair of summaries receives a low ROUGE or BLEU score, neither are necessarily incorrect. To illustrate this point, Fig. 3 shows a consumer health question, two summaries written by different annotators, and the ROUGE-2 score between the summaries. While the ROUGE-2 score is quite low, both summaries contain information relevant to the question, the main difference between the two summaries being that one focuses on genetics, the other on family history. Though the metrics certainly are useful for measuring similarity, there is difficulty in quantitatively measuring the differences between summaries such as these. Fortunately, the development of metrics for this purpose is an active area of research[14].

Experimental benchmarking. To benchmark the MEDIQA-AnS dataset and demonstrate how it can be used to evaluate automatically generated summaries, we conducted a series of experiments using a variety of summarization approaches. Three baseline and three deep learning algorithms were implemented and are listed below:

Lead-3. The Lead-3 baseline takes the first three sentences of an article as a summary. This has been shown in previous work[15,16] to be an effective baseline for summarization.

k random sentences. Similarly to the Lead-3 baseline, we select $k=3$ random sentences from each article.

k-best ROUGE. We select $k=3$ sentences with the highest ROUGE-L score relative to the question.

Bidirectional long short-term memory (BiLSTM) network. A BiLSTM[17] was trained to select the most relevant sentences in an article, similar to other extractive LSTM models[18,19].

Pointer-generator network. The Pointer-Generator network[15] is a hybrid sequence-to-sequence attentional model, with the capability to create summaries by copying text from the source document while also generating novel text.

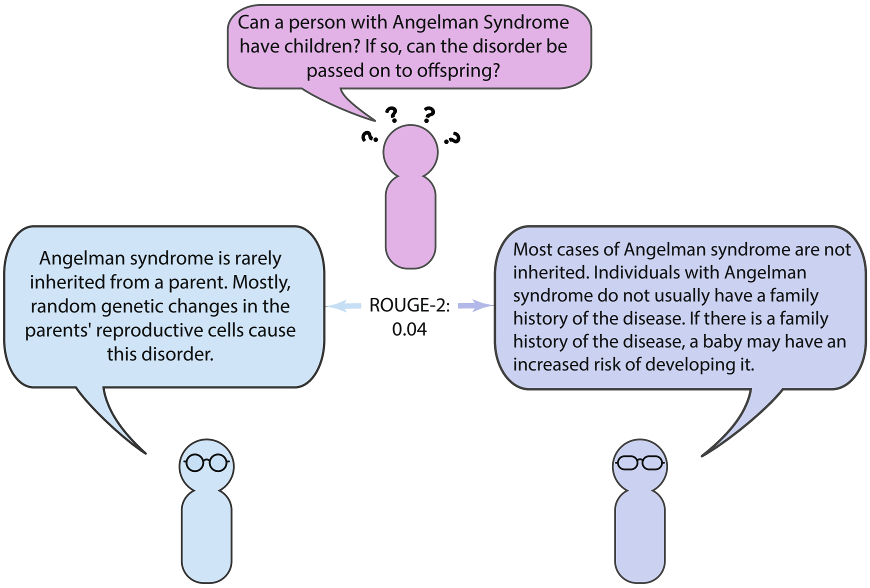

Can a person with Angelman Syndrome have children? If so, can the disorder be passed on to offspring?
Angelman syndrome is rarely inherited from a parent. Mostly, random genetic changes in the parents' reproductive cells cause this disorder.
ROUGE-2: 0.04
Most cases of Angelman syndrome are not inherited. Individuals with Angelman syndrome do not usually have a family history of the disease. If there is a family history of the disease, a baby may have an increased risk of developing it.

Fig. 3 Comparison of summaries of an answer to a consumer health question. While both summaries are relevant, they receive a low ROUGE-2 score.

Experiment	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
Lead-3	0.23	0.11	0.08	0.04
3-random	0.20	0.08	0.06	0.04
3-best ROUGE	0.16	0.08	0.06	0.00
BiLSTM	0.22	0.10	0.08	0.03
Pointer-Generator	0.21	0.09	0.07	0.03
BART	0.29	0.15	0.12	0.09

Table 6. Automatically generated summaries compared to extractive summaries.

Bidirectional autoregressive transformer (BART). BART[20] is a recently published transformer-based encoder-decoder model combining a bidrectional encoder similar to BERT[21] and an auto-regressive decoder similar to GPT-2[22]. To improve performance, instead of training the model directly on data relevant for summarization and other language-generation tasks, the authors first pre-trained the model on objectives designed to improve its general ability to understand the content of text. These objectives include document rotation, sentence permutation, text-infilling, token masking and token deletion. Given text which has been corrupted by one of these methods, the model is asked to de-noise the document or sequence. This pre-training procedure allows BART to generate text of higher quality when it is later fine-tuned on a more task-specific dataset, such as one for summarization. Pre-training approaches have been shown to achieve state-of-the-art results on a wide variety of NLP tasks, summarization included[20–22].

In our experiments, all machine learning models were trained using the questions, abstracts, and snippets available in the BioASQ data, which can be easily adapted for training summarization models. Essentially, we treat the scientific abstracts in the collection as the source documents, and the snippets extracted by the creators of the collection as the summaries. The snippets provide information relevant to answering the questions, which are biomedical in domain; for example, *Is Hirschsprung disease a mendelian or a multifactorial disorder?* To compute validation loss during training, we used the medical question and answer dataset MedInfo[23]. This dataset consists of answers selected from reliable online health information, in response to consumer health questions about medications. It is therefore similar in structure and content to the MEDIQA-QA data, and can be used to approximate the single document, extractive summarization task provided in the MEDIQA-AnS collection.

We use these methods to automatically summarize the full text of the web pages in MEDIQA-AnS. Tables 6 and 7 show the comparison between the automatically generated summaries and the manually generated summaries. We include results for only single document summarization; however, the same experiments could be run in a multi-document setting.

Generated examples from Lead-3, the Pointer-Generator, and BART can be seen in Table 8. These show that the quality of the source text in MEDIQA-Ans is suitable for use with generative deep learning models. The text of BART and the Pointer-Generator is grammatical and, particularly for BART, relevant to the question asked. In terms of ROUGE and BLEU, the Lead-3 baseline compares well to the other methods, as expected. Relevant to machine learning research, BART outperforms the Pointer-Generator ($p < 0.0039$, across all experiments,

Experiment	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
Lead-3	0.25	0.10	0.07	0.06
3-random	0.22	0.07	0.04	0.05
3-best ROUGE	0.17	0.07	0.04	0.02
BiLSTM	0.24	0.09	0.06	0.06
Pointer-Generator	0.24	0.08	0.06	0.05
BART	0.32	0.12	0.08	0.09

Table 7. Automatically generated summaries compared to abstractive summaries.

Model	Generated text
Question:	*what is copd*
Lead-3	COPD What is COPD? Chronic obstructive pulmonary disease, or COPD, is a progressive lung disease in which the airways of the lungs become damaged, making it hard to breathe.
Pointer-Generator	Overview Chronic obstructive pulmonary disease (COPD) is a chronic inflammatory lung disease that causes obstructed airflow from the lungs. Symptoms include breathing difficulty, cough, mucus (sputum) production and wheezing. It's caused by long-term exposure to irritating gases or particulate matter, most often from cigarette smoke. People with COPD are at increased risk of developing heart disease, lung cancer and a variety of other conditions. Emphysema and chronic bronchitis are the two most common conditions that contribute to COPD.
BART	Chronic obstructive pulmonary disease, or COPD, is a progressive lung disease in which the airways of the lungs become damaged, making it hard to breathe. You may also have heard COPD called other names, like emphysema or chronic bronchitis.
Question:	*How bad can endometriosis get?*
Lead-3	Endometriosis occurs when cells from the lining of your womb (uterus) grow in other areas of your body. This can cause pain, heavy bleeding, bleeding between periods, and problems getting pregnant (infertility). Causes Every month, a woman's ovaries produce hormones that tell the cells lining the uterus to swell and get thicker
Pointer-Generator	This growths stay in your body, and like the cells in the lining of your uterus, these growths react to the hormones from your ovaries. They grow and bleed when you get your period. Over time, the growths may add more tissue and blood.
BART	Endometriosis occurs when cells from the lining of your womb (uterus) grow in other areas of your body. This can cause pain, heavy bleeding, bleeding between periods, and problems getting pregnant (infertility). Causes Every month, a woman's ovaries produce hormones that tell the cells lining the uterus to swell and get thicker. Your uterus sheds these cells along with blood and tissue through your vagina when you have your period.

Table 8. Examples of summaries generated by Lead-3, Pointer-Generator, and BART, in response to consumer-health questions.

Experiment	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
Pages				
BART + Q, Abstractive	0.32	0.12	0.08	0.09
BART, Abstractive	0.26	0.09	0.05	0.07
BART + Q, Extractive	0.29	0.15	0.12	0.09
BART, Extractive	0.24	0.10	0.07	0.05
Passages				
BART + Q, Abstractive	0.46	0.29	0.24	0.19
BART, Abstractive	0.43	0.27	0.21	0.17
BART + Q, Extractive	0.46	0.37	0.35	0.18
BART, Extractive	0.43	0.35	0.33	0.14

Table 9. BART using question-driven approach. Shows summaries generated with and without access to the question, compared to extractive and abstractive summaries. Across all experiments, BART scores higher when the question (+Q) is included in the input.

Wilcoxon signed-rank test), which is consistent with previous work[20]. Additionally, it is interesting to note that the ROUGE-1 and BLEU scores for each model tend to increase in the abstractive evaluation, and the ROUGE-2 and ROUGE-L scores tend to increase in the extractive evaluation. The increase in ROUGE-1 is potentially due to the fact that, as indicated in Table 3, the abstractive summaries are shorter, which makes it easier for an automatically generated summary to contain a larger percentage of unigrams present in the reference summary. ROUGE-2 and ROUGE-L may increase because it is easier for a system to get a sequence of words correct in an extractive setting. These observations may be useful to consider while using MEDIQA-AnS, since the type of summary and the metric used for evaluation will affect the observed performance of a system.

To determine if the dataset could be used for the evaluation of question-driven summarization, we trained BART with and without access to the consumer health questions. For training and testing in these experiments, we concatenated the question to the beginning of the article. This approach is similar to other deep learning text generation work, where including unique text at the beginning of the documents fed to a model can give greater control over the content of the output. For example, users can provide the CTRL model[24] with control codes to specify the topic of generated text.

After including the question with the text during training and evaluation, we observed a significant difference between the two types of summaries. Table 9 shows that including the question with the input documents significantly improved BART's performance across all summarization tasks: passages and pages, extractive and abstractive ($p < 0.0064$ across all experiments, Wilcoxon signed-rank test). This indicates that being able to compare plain summaries with question-driven summaries is important to developing medical question answering systems, and since there are no existing datasets for this type of evaluation, MEDIQA-AnS can play a valuable role for further research in this area.

Usage Notes

We have provided instructions in the README file in the Open Science Framework repository describing how to process the MEDIQA-AnS dataset. Examples of processing the data for different summarization evaluations can be found in the code located at the GitHub repository provided below.

Code availability

The code to process MEDIQA-AnS and reproduce the results of the experimental benchmarks shown here can be found at https://www.github.com/saverymax/qdriven-chiqa-summarization.

Received: 17 March 2020; Accepted: 28 August 2020;
Published online: 02 October 2020

References

1. Mishra, R. *et al*. Text summarization in the biomedical domain: a systematic review of recent research. *Journal of Biomedical Informatics* **52**, 457–467 (2014).
2. Demner-Fushman, D., Mrabet, Y. & Ben Abacha, A. Consumer health information and question answering: helping consumers find answers to their health-related information needs. *Journal of the American Medical Informatics Association* **27**, 194–201 (2019).
3. National Network of Libraries of Medicine. The consumer health reference interview and ethical issues. *National Network of Libraries of Medicine Initiatives*, https://nnlm.gov/initiatives/topics/ethics (2020).
4. Higgins, J. P. T. *et al*. Cochrane Handbook for Systematic Reviews of Interventions, version 6.0. *Cochrane* www.training.cochrane.org/handbook (2019).
5. Ben Abacha, A. & Demner-Fushman, D. On the Summarization of Consumer Health Questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 2228–2234. Association for Computation Linguistics, Florence, Italy, 2019).
6. Savery, M., Ben Abacha, A., Gayen, S. & Demner-Fushman, D. Questiondriven summarization of answers to consumer health questions. *Open Science Framework* https://doi.org/10.17605/OSF.IO/FYG46 (2020).
7. Hermann, K. M. *et al*. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28*, 1693–1701 (Curran Associates, Inc., 2015).
8. Cohan, A. *et al*. 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* (pp. 615–621. Association for Computation Linguistics, New Orleans, Louisiana, 2018).
9. Fabbri, A. R., Li, I., She, T., Li, S. & Radev, D. R. Multi-news: a largescale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1074–1084. Association for Computation Linguistics, Florence, Italy, 2019).
10. Tsatsaronis, G. *et al*. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics* **16**, 138 (2015).
11. Ben Abacha, A., Shivade, C. & Demner-Fushman, D. Overview of the mediqa 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task* (pp. 370–379. Association for Computation Linguistics, Florence, Italy, 2019).
12. Lin, C.-Y. ROUGE: a package for automatic evaluation of summaries. In *Text Summarization Branches Out* (pp. 74–81. Association for Computational Linguistics, Barcelona, Spain, 2004).
13. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318. Association for Computational Linguistics, Philadelphia, USA, 2002).
14. Wang, A., Cho, K. & Lewis, M. Asking and Answering Questions to Evaluate the Factual Consistency of Summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5008–5020. Association for Computation Linguistics, Online, 2020).
15. See, A., Liu, P. J. & Manning, C. D. Get to the point: summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (pp. 1073–1083. Association for Computation Linguistics, Vancouver, Canada, 2017).
16. Nallapati, R., Zhai, F. & Zhou, B. Summarunner: a recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (pp. 3075–3081. AAAI Press, San Francisco, USA, 2017).
17. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural computation* **9**, 1735–80 (1997).
18. Chen, Y.-C. & Bansal, M. Fast abstractive summarization with reinforceselected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (pp. 675–686. Association for Computation Linguistics, Melbourne, Australia, 2019).
19. Liu, Y. & Lapata, M. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5070–5081 (Association for Computational Linguistics, Florence, Italy, 2019).
20. Lewis, M. *et al*. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7871–7880. Association for Computational Linguistics, Online, 2020).
21. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 4171–4186. Association for Computation Linguistics, Minneapolis, Minnesota, 2019).

22. Radford, A. *et al*. Language models are unsupervised multitask learners. *OpenAI Blog* https://openai.com/blog/better-language-models (2019).
23. Ben Abacha, A. *et al*. Bridging the gap between consumers' medication questions and trusted answers. In *MEDINFO 2019: Health and Wellbeing e-Networks for All*, 25–29 (IOS Press, 2019).
24. Keskar, N. S., McCann, B., Varshney, L., Xiong, C. & Socher, R. CTRL - A Conditional Transformer Language Model for Controllable Generation. Preprint at https://arxiv.org/abs/1909.05858 (2019).

Acknowledgements

This research was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine, and Lister Hill National Center for Biomedical Communications.

Author contributions

M.S. implemented data processing code and pipelines, conducted baseline and machine learning experiments, contributed to generating the manual summaries, and wrote and edited the manuscript. A.B.-A. authored the MEDIQA data used as the backbone for the collection presented here, as well as the MedInfo data used for training, provided guidance on their use, developed the summarization interface, edited the manuscript, and provided feedback for collection creation. S.G. managed the interface for generating the summaries, and provided data processing support. D.D.-F. conceived of the project, contributed to generating the manual summaries, edited the manuscript, and otherwise provided feedback on all aspects of the study.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.D.-F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.