An exploratory study of abstractive text summarization using a sequence-to-sequence model

M.Kavitha [1], Research Scholar, Dr.K.Akila [2], Assistant Professor
[1,2]Department of CSE, SRM Institute of Science and Technology, College of Engineering and Technology,
Vadapalani, Chennai, India
[1] km2693@srmist.edu.in, [2] akilak@srmist.edu.in

2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS) | 979-8-3503-9458-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCEBS58601.2023.10448613

Abstract—**Text summarization has evolved over a period of time in various domains and benefits most professionals and researchers. To provide salient summarization in a short span of time, various approaches to text summarization are discovered. The objective of the paper is to increase the efficiency of the model by delivering a concise and precise summary. The intention of text summarization is to provide sufficient topic coverage and readability. The paper also presents the applications and limitations of text summarization. The study utilized a sequence-to-sequence model stacked with LSTM networks for text summarization. The results of the sequence-to-sequence model are better than other conventional models.**

Keywords—Abstractive summarization, Decoder, Encoder, Extractive summarization, LSTM, Sequence-to-Sequence model

I. Introduction

Automatic text summarization [1-4] aims to generate shortened versions from lengthy documents which is something difficult and costly to undertake manually. Text summarization is useful when the volume of content is large. In text summarization, there are two primary approaches: abstractive summarization and extractive summarization. Extractive summarization is easy when compared to abstractive summarization. It takes the text from the original document.

Abstractive summarization [5,6] is quite hard and generates novel sentences. Abstractive summarization rewrites and reformulates the text. It is similar to human summarization. It works on critical information of the original text and generates new text. This approach entails extracting the key information, understanding the context, and constructing new text. Abstractive summarization is more complex when compared to extractive summarization as it requires extracting relevant information as well as recreating new text. It works well with deep learning models.

Word and phrase frequency methods can also be used for automatic summarization. The weight of sentences is calculated using the cue method, title method, and location method. Extractive summarization can be done by intermediate representation, sentiment score, and sentence selection. Topic words, frequency-based approach, word probability, TFIDF, centroid-based summarization, LSA, and Bayesian topic model are the approaches to topic representation. Web summarization, scientific article summarization, and email summarization are the ways to do content summarization. Graph method and machine learning are the methods for text summarization.

II. Background and Significance of the study

The authors [7] Altmami, Nouf Ibrahim, and Mohamed El Bachir Menai performed a survey on automatic summarization of scientific articles. Either single or multiple articles can be summarized. The authors suggest conceptual frameworks for abstract-based and citation-based article summarization in this paper.

Han, Yi, Gaurav Nanda, and Mohsen Moghaddam [8] proposed a framework for automated generated summaries from online reviews with sentiment and attributes. The authors suggested optimizing the opinion summarization framework to improve the ROUGE score and reduce the loss function.

Esteva, Andre, et al. [9] the authors developed a query-based search using Siamese BERT and two keyword-based search models using TF-IDF and BM25 to retrieve information related to COVID. The documents are assigned relevance scores to the documents from the question answering and abstractive summarization.

Searle, Thomas, et al. [10] proposed an architecture for abstractive and extractive summarization of inpatient documentation. It is quite a complex process to summarize medical documentation as it contains medical terminology. In this study, GloVe and S-BERT are used for embedding and computing sentence averages. Text rank is used to retrieve the foremost N sentences from the input document. Researchers from the healthcare domain quickly grasp the findings from the clinical notes with the help of automatic summarization.

Qiu, Yunjian, and Yan Jin. [11] proposed extractive summarization using word embedding generated through BERT and k-means clustering techniques. In this work, a systematic method is suggested for capturing domain knowledge and generating summarizations. Summaries with various dimensions are generated using sentence embedding and Rouge scores computed from BERT and non-BERT models.

979-8-3503-9458-0/23/$31.00 ©2023 IEEE

Authors Anand, Deepa, and Rupali Wagh [12] employ a pair of architectures for word and sentence embedding to grasp the textual semantics. Notably, their proposed approach operates independently of domain-specific knowledge. In this research, similarity measurements are derived from the intersection of crucial terms, TF-IDF scores, Rouge-L scores, and sentence embeddings (SSE). The summary is generated using FFNN and LSTM architectures. Rouge-1 and Rouge-L scores are observed by varying the summary size of the document.

Previous research has been done using TFIDF, textrank, BERT, embeddings, DNN, and clustering techniques. So, this study performs abstractive text summarization on a single document using encoder and decoder techniques. It uses a news summary article for summarization.

III. Materials and Methods

Automatic summarization reduces users reading time. It helps in selecting the documents for various purposes (research). Automatic summarization can be utilized for effective indexing. It exhibits a lower degree of bias compared to human summarization. Personalized summaries prove valuable in question-answering systems, as they furnish tailored information. The personalized summary must contain the primary attributes of the article. The generated summary should increase the relevance score and reduce the redundant content from the documents.

A. Types of Summarization

Considering the input, text summarization can be classified into a single document or multiple documents [13]. As for the output, text summarization falls into extractive summarization and abstractive summarization categories. Other types of summarization techniques are generic summarization, domain-specific summarization, and query-based summarization [14,15]. Various text summarization approaches are shown in Figure 1.

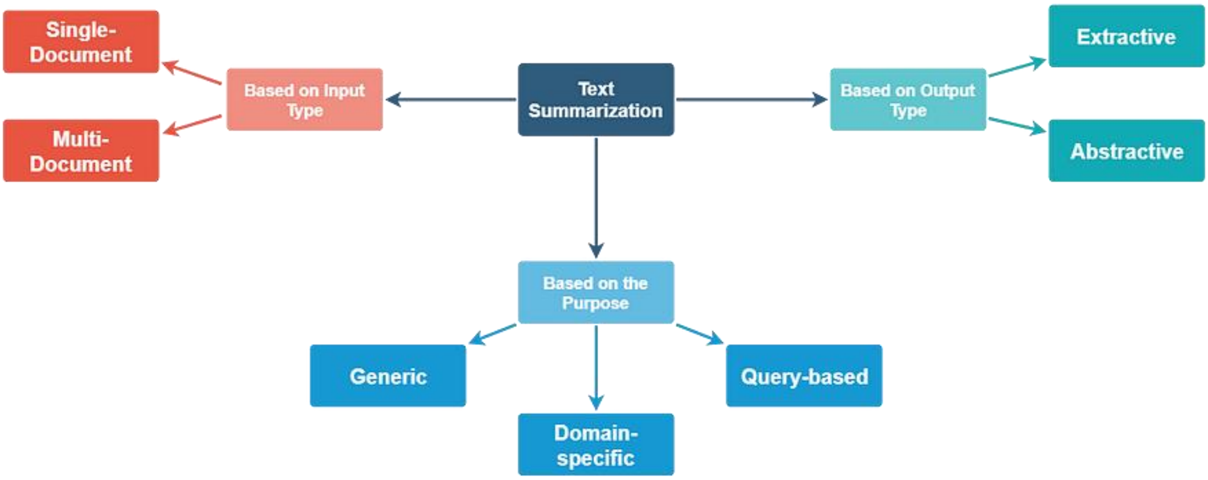


Fig. 1. Summarization Types

B. Feature Extraction

Summarization can be done by extracting significant terms or key phrases from the original document. To identify such terms, techniques like word sequences, linguistic patterns, or part-of-speech tagging can be utilized. Positive and negative labeled keyphrases can also be used to extract important terms. The length and frequency of the keyphrases, the word occurrence in keyphrases, and the number of characters in the keyphrases are some of the features of text summarization.

C. Summarization using TFIDF

To perform text summarization using TFIDF, the entire document is split into sentences. Then score for each sentence is computed using the TFIDF matrix. It is a sentence versus terms matrix. As TFIDF is a sparse matrix, the evaluation score is computed by taking an average of non-zero TFIDF values. The summation value will be biased towards longer sentences hence the average score is taken. The sentences with the highest score are selected for summarization.

D. Summarization using TextRank

The summary from TF-IDF is much more precise and short as compared to the one by GenSim Text Rank. TF-IDF summary does not have any punctuations and digits, it is pure text with the words that have the maximum frequency in the original text. GenSim Text Rank gives a summary longer than TF-IDF. Also, it includes \n which indicates newlines, includes digits, and punctuations like brackets which are not in weighted frequency, and TF-IDF summary. The weighted frequency of the words is taken to build a summary while in GenSim text Rank it does not take the weighted frequency of words into consideration for summarizing. In TF-IDF we perform tokenization and remove stop words, punctuations, and use word frequencies to build the summaries

E. Summarization using N-grams

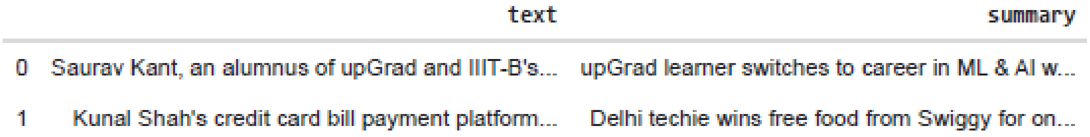
Text summarization can be improved using N-gram models with the NLTK package as it gives precise summarization compared to the word frequency, TD-IDF, and GenSim Text Rank summarization. In word frequency summarization, digits and punctuations are included in the summary. It contains the most frequently used words and the summary is big. When using N-grams, the summary is small as it does not include digits, brackets, and punctuation. It uses the probability of the words and word prediction to generate the summary. The traditional summarization methods classify the sentences based on the word frequency, position in the content, and cue expressions. N-gram models are better than these traditional methods.

F. Summarization using Seq2Seq model

Sequence-to-Sequence (Seq2Seq) models have gained widespread popularity in the realm of Natural Language Processing (NLP), with their remarkable capability to effectively manage sequences of varying lengths in both input and output. The architecture of a Seq2Seq model primarily comprises two pivotal components: the Encoder and the Decoder. These components are essential in processing and generating sequences of tokens. In the training phase, Seq2Seq models are fed with pairs of input and output sequences. The encoder and decoder use either LSTM or GRU. It does not use RNN because of the vanishing gradient problem. Special tokens "start" and "end," are thoughtfully introduced into the target sequence. The decoder recognizes the sentence with the help of these special tokens. Through training, the model endeavors to maximize the likelihood of generating the correct output sequence, given a specific input sequence as its context. Seq2Seq models are flexible and it is capable of capturing the context. It easily handles sequential data and using the attention mechanism it can emphasize specific parts of the input text.

IV. Results and Discussion

The news_summary is the dataset utilized for this study to perform summarization. It comprises a total of 98,353 samples, with 88,517 samples allocated for training and the remaining 9,836 samples designated for testing. The experiment uses the sequence-to-sequence model that contains an encoder and decoder. Both the encoder and decoder contain the LSTM network. Encoder and decoder are used in the seq2seq model for converting text sequence to integer sequence and vice versa.



	text	summary
0	Saurav Kant, an alumnus of upGrad and IIIT-B's...	upGrad learner switches to career in ML & Al w...
1	Kunal Shah's credit card bill payment platform...	Delhi techie wins free food from Swiggy for on...

Fig. 2. Samples of news-summary files before data cleansing.

The first two samples of the news_summary file before data cleansing are shown in Figure 2. The order of regular expression is important while preprocessing the text. In data cleansing, escape sequences, digits, special characters, extra spaces, single characters between two spaces, and URLs are removed. The approximate time to clean the text column and summary column is 7 minutes and 1 minute. The start and end tags are added to the summary column of the news article. The average text length is between 0 to 70 words and the average summary length is between 0 to 15 words. The percentage of summaries having 0-15 words and the percentage of summaries having 0-70 words are depicted in Figure 3. The source text after data preparation is displayed in Figure 4.

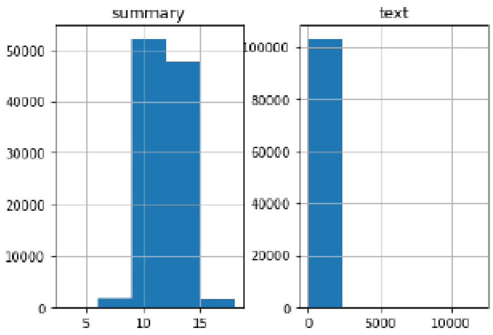
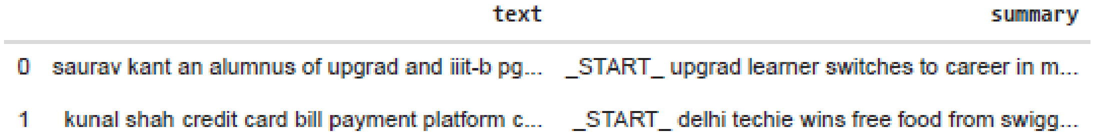


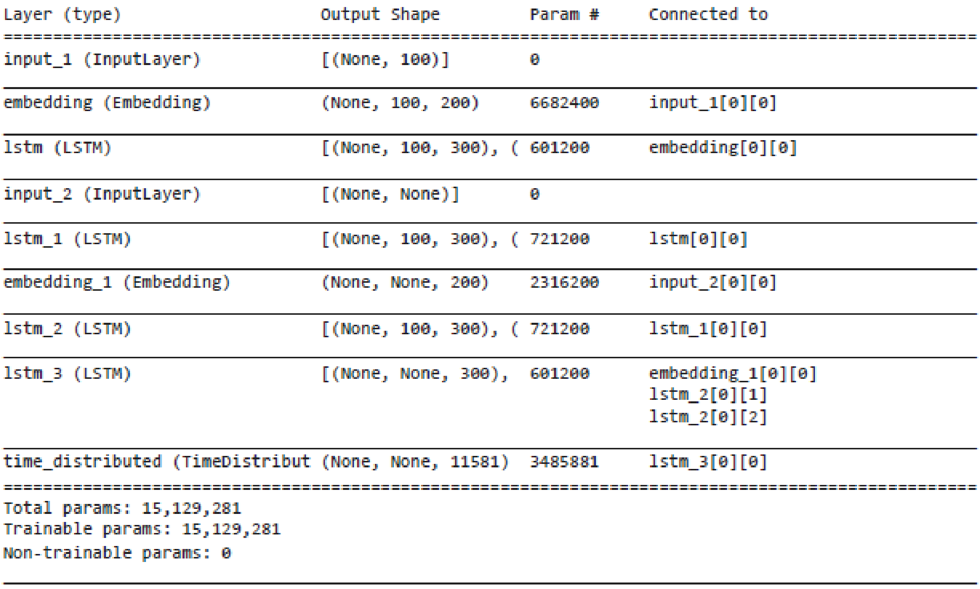
Fig. 3. The average length of text and summary feature.



	text	summary
0	saurav kant an alumnus of upgrad and iiit-b pg...	_START_ upgrad learner switches to career in m...
1	kunal shah credit card bill payment platform c...	_START_ delhi techie wins free food from swigg...

Fig. 4. Samples after data cleansing.

Tokenizing the text using the tokenizer function. The text sequences are converted into integer sequences using the encoder. Later padding the text to a maximum length that is fixed to perform embedding operation. Word2vec embedding technique is applied for this model. Unique words in the text, rare words whose count is below the threshold value, and the most common words are computed for data analysis. Common words are found by finding the difference between the number of unique words and rare words. The size of the vocabulary is 33412. The sequence-to-sequence model summary is displayed in Figure 5. Rmsprop Optimizer and sparse categorical cross entropy are used during compilation. The model trains 88517 observations and validates 9836 observations in 50 epochs. The loss curve for training and validation is depicted in Figure 6. Figure 7 shows the values of the review field, original summary, and predicted summary.



```
Layer (type)                    Output Shape         Param #     Connected to
==================================================================================================
input_1 (InputLayer)            [(None, 100)]        0
__________________________________________________________________________________________________
embedding (Embedding)           (None, 100, 200)     6682400     input_1[0][0]
__________________________________________________________________________________________________
lstm (LSTM)                     [(None, 100, 300), ( 601200      embedding[0][0]
__________________________________________________________________________________________________
input_2 (InputLayer)            [(None, None)]       0
__________________________________________________________________________________________________
lstm_1 (LSTM)                   [(None, 100, 300), ( 721200      lstm[0][0]
__________________________________________________________________________________________________
embedding_1 (Embedding)         (None, None, 200)    2316200     input_2[0][0]
__________________________________________________________________________________________________
lstm_2 (LSTM)                   [(None, 100, 300), ( 721200      lstm_1[0][0]
__________________________________________________________________________________________________
lstm_3 (LSTM)                   [(None, None, 300),  601200      embedding_1[0][0]
                                                                 lstm_2[0][1]
                                                                 lstm_2[0][2]
__________________________________________________________________________________________________
time_distributed (TimeDistribut (None, None, 11581)  3485881     lstm_3[0][0]
==================================================================================================
Total params: 15,129,281
Trainable params: 15,129,281
Non-trainable params: 0
__________________________________________________________________________________________________
```

Fig. 5. Sequence to sequence model summary.

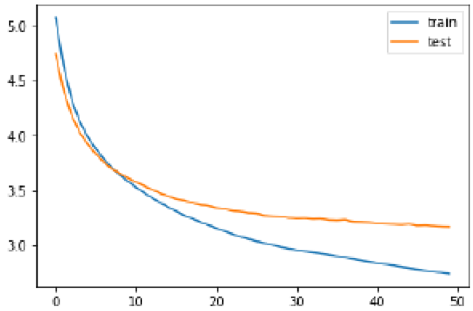
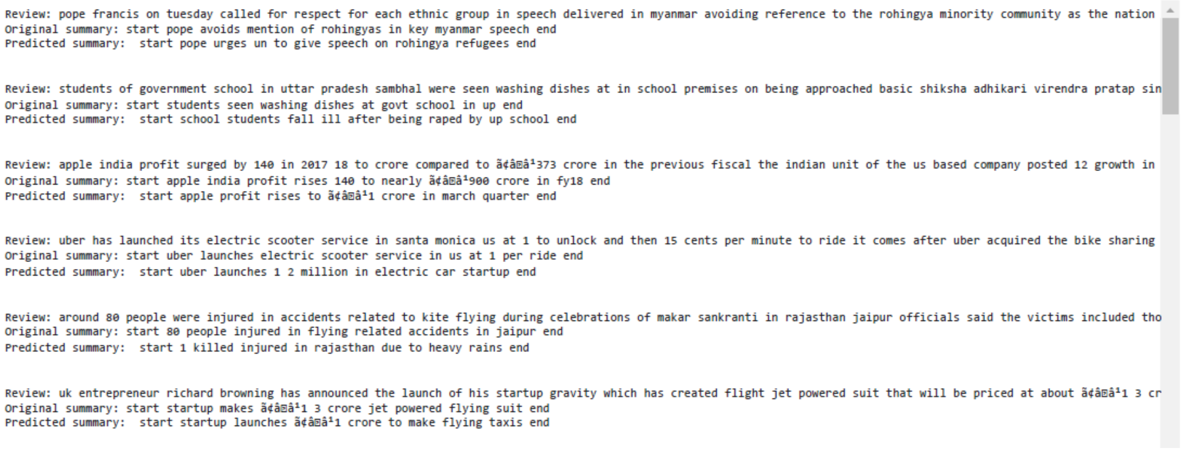


Fig. 6. The loss curve for training and testing.



```
Review: pope francis on tuesday called for respect for each ethnic group in speech delivered in myanmar avoiding reference to the rohingya minority community as the nation
Original summary: start pope avoids mention of rohingyas in key myanmar speech end
Predicted summary:  start pope urges un to give speech on rohingya refugees end


Review: students of government school in uttar pradesh sambhal were seen washing dishes at in school premises on being approached basic shiksha adhikari virendra pratap sin
Original summary: start students seen washing dishes at govt school in up end
Predicted summary:  start school students fall ill after being raped by up school end


Review: apple india profit surged by 140 in 2017 18 to crore compared to ã¢â□â¹373 crore in the previous fiscal the indian unit of the us based company posted 12 growth in
Original summary: start apple india profit rises 140 to nearly ã¢â□â¹900 crore in fy18 end
Predicted summary:  start apple profit rises to ã¢â□â¹1 crore in march quarter end


Review: uber has launched its electric scooter service in santa monica us at 1 to unlock and then 15 cents per minute to ride it comes after uber acquired the bike sharing
Original summary: start uber launches electric scooter service in us at 1 per ride end
Predicted summary:  start uber launches 1 2 million in electric car startup end


Review: around 80 people were injured in accidents related to kite flying during celebrations of makar sankranti in rajasthan jaipur officials said the victims included tho
Original summary: start 80 people injured in flying related accidents in jaipur end
Predicted summary:  start 1 killed injured in rajasthan due to heavy rains end


Review: uk entrepreneur richard browning has announced the launch of his startup gravity which has created flight jet powered suit that will be priced at about ã¢â□â¹1 3 cr
Original summary: start startup makes ã¢â□â¹1 3 crore jet powered flying suit end
Predicted summary:  start startup launches ã¢â□â¹1 crore to make flying taxis end
```

Fig. 7. The values of the review field, original summary, and predicted summary.

A. Limitations of text summarization

Some of the challenges in text summarization are identifying key topics in the context, and interpreting, generating summaries, and evaluating summaries. Sufficient tools and techniques are necessary for domain-specific text summarization. The end results of text summarization must produce a concise and accurate summary. An appropriate feature analysis or topic modeling technique is required to extract relevant key phrases from the document. With the help of NLP, machine learning, and deep learning techniques the semantics of the context can be learned. The similarity of sentences and sentence scores can be utilized to generate the summary. Various metrics are utilized to assess the summary-generated model performance.

V. Conclusion

Text summarization has evolved over a period of time in various domains and benefits most professionals and researchers. To provide salient summarization in a short span of time, various approaches to text summarization are discovered. The aim of this article is to apply the sequence-to-sequence model to generate a concise and precise summary. The sequence-to-sequence model is used for several NLP applications. It provides good results when compared to traditional methods. In the future, the summarization can be implemented by stacking GRUs instead of LSTM networks. The performance can also be assessed by including attention layers that focus on specific parts of the input text.

References

[1] Yogesh Kumar, Meena, and Dinesh Gopalani., Procedia Computer Science, Domain independent framework for automatic text summarization, 48 (2015): pp 722-727.

[2] Deutsch, Rotem Dror, Daniel, and Dan Roth., Transactions of the Association for Computational Linguistics 9 (2021): pp 1132-1146, A statistical analysis of summarization evaluation metrics using resampling methods.

[3] Rohil, Mukesh Kumar, and Varun Magotra., Healthcare Analytics 2 (2022): 100058, An exploratory study of automatic text summarization in biomedical and healthcare domain.

[4] Muniraj, K. R. Sabarmathi, Padhma, and R. Leelavathi., International Journal of Intelligent Networks 4 (2023): pp 53-61, HNTSumm: Hybrid text summarization of transliterated news articles,

[5] Wanjale, Kirti, et al., International Journal of Engineering Research and Technology (IJERT) ISSN: 2278-0181, Comprehensive Survey on Abstractive Text Summarization.

[6] Rahman, Md Motiur, and Fazlul Hasan Siddiqui., Etri Journal, Multi-layered attentional peephole convolutional LSTM for abstractive text summarization, 43.2 (2021): pp 288-298.

[7] Altmami, Nouf Ibrahim, and Mohamed El Bachir Menai., Journal of King Saud University-Computer and Information Sciences 34.4 (2022):1011-1028, Automatic summarization of scientific articles: A survey.

[8] Han, Yi, Gaurav Nanda, and Mohsen Moghaddam., Journal of Mechanical Design 145.4 (2023): 041401, Attribute-Sentiment-Guided Summarization of User Opinions From Online Reviews.

[9] Esteva, Andre, et al., NPJ digital medicine 4.1 (2021): 68, COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization.

[10] Searle, Thomas, et al., Journal of Biomedical Informatics 141 (2023): 104358, Discharge summary hospital course summarisation of in patient Electronic Health Record text with clinical concept guided deep pre-trained Transformer models.

[11] Qiu, Yunjian, and Yan Jin., Journal of Computing and Information Science in Engineering, Engineering Document Summarization: A Bidirectional Language Model-Based Approach, 22.6 (2022): 061004.

[12] Anand, Deepa, and Rupali Wagh., Journal of King Saud University-Computer and Information Sciences 34.5 (2022):2141-2150, Effective deep learning approaches for summarization of legal texts,

[13] K., PV Venkateswara Rao, Veningston, and M. Ronalda, Procedia Computer Science 218 (2023): pp 1220-1228, Personalized Multi-document Text Summarization using Deep Learning Techniques.

[14] Litvak, Marina, and Natalia Vanetik. Proceedings of the multiling 2017 workshop on summarization and summary evaluation across source types and genres, Query-based summarization using MDL principle.

[15] Savery, Max, et al. Scientific Data 7.1 (2020): 322. "Question-driven summarization of answers to consumer health questions."