frontiers in Digital Health

BRIEF RESEARCH REPORT
published: 04 December 2020
doi: 10.3389/fdgth.2020.585559

A Light-Weight Text Summarization System for Fast Access to Medical Evidence

Abeed Sarker[1,2]*, *Yuan-Chi Yang*[1], *Mohammed Ali Al-Garadi*[1] *and Aamir Abbas*[3]

[1] *Department of Biomedical Informatics, School of Medicine, Emory University, Atlanta, GA, United States,* [2] *Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, United States,* [3] *Heinz College of Information Systems and Public Policy, Carnegie Mellon University, Pittsburgh, PA, United States*

As the volume of published medical research continues to grow rapidly, staying up-to-date with the best-available research evidence regarding specific topics is becoming an increasingly challenging problem for medical experts and researchers. The current COVID19 pandemic is a good example of a topic on which research evidence is rapidly evolving. Automatic query-focused text summarization approaches may help researchers to swiftly review research evidence by presenting salient and query-relevant information from newly-published articles in a condensed manner. Typical medical text summarization approaches require domain knowledge, and the performances of such systems rely on resource-heavy medical domain-specific knowledge sources and pre-processing methods (e.g., text classification) for deriving semantic information. Consequently, these systems are often difficult to speedily customize, extend, or deploy in low-resource settings, and they are often operationally slow. In this paper, we propose a fast and simple extractive summarization approach that can be easily deployed and run, and may thus aid medical experts and researchers obtain fast access to the latest research evidence. At runtime, our system utilizes similarity measurements derived from pre-trained medical domain-specific word embeddings in addition to simple features, rather than computationally-expensive pre-processing and resource-heavy knowledge bases. Automatic evaluation using ROUGE—a summary evaluation tool—on a public dataset for evidence-based medicine shows that our system's performance, despite the simple implementation, is statistically comparable with the state-of-the-art. Extrinsic manual evaluation based on recently-released COVID19 articles demonstrates that the summarizer performance is close to human agreement, which is generally low, for extractive summarization.

Keywords: medical text processing, text summarization, text mining, natural language processing, health informatics, extractive summarization

OPEN ACCESS

Edited by:
Hercules K. Dalianis,
Stockholm University, Sweden

Reviewed by:
Hans Moen,
University of Turku, Finland
Rebecka Weegar,
Stockholm University, Sweden

****Correspondence:***
Abeed Sarker
abeed@dbmi.emory.edu

Specialty section:
This article was submitted to
Health Informatics,
a section of the journal
Frontiers in Digital Health

Received: *21 July 2020*
Accepted: *13 November 2020*
Published: *04 December 2020*

Citation:
Sarker A, Yang Y-C, Al-Garadi MA and Abbas A (2020) A Light-Weight Text Summarization System for Fast Access to Medical Evidence. Front. Digit. Health 2:585559. doi: 10.3389/fdgth.2020.585559

INTRODUCTION

The overarching objective of evidence-based medicine practice is to actively incorporate the best available and most reliable scientific evidence into clinical practice guidelines and decision-making (1). The movement associated with the establishment of evidence-based medicine practice has led to the development of evidence hierarchies for medical research, establishment of clinical practice guidelines, and recognition of the importance of patient-oriented evidence (2, 3). Since the inception of the modern concept of evidence-based medicine, medical practitioners have been advised to combine their clinical expertise and understanding of patients' priorities with the latest scientific evidence (4–6). Early and recent studies have extensively discussed the problem of *information overload* that many practitioners face, particularly in clinical settings, due to the massive amounts of research evidence that is available and the continuous growth of such evidence (7). Searching through medical evidence regarding a specific topic is time-consuming, and practitioners often consider the task to be unproductive and futile (8–10). PubMed[1], which indexes over 30 million articles, typically returns multiple pages of research publications even when the queries are very targeted and specific. Almost two decades ago, Hersh et al. (11) discussed the long time (30 min, on average) that it takes for experienced practitioners to search for evidence, and, particularly at point-of-care, practitioners cannot afford to spend that much time. Over time, with the increasing rate of publication of medical literature, these problems associated with evidence curation have only increased (12). Improved literature searching and fast access to relevant and summarized information can be particularly beneficial for medical students and young practitioners because of their lack of clinical experience, or at times when there is a burst of growth in research evidence on a topic (e.g., the ongoing COVID19 pandemic).

Natural language processing (NLP) and information retrieval methods have the potential to aid medical experts and researchers to collect and review the latest and emerging research evidence in an efficient manner. NLP methods can, for example, help experts formulate effective search queries and summarize individual publications. Query-focused text summarization approaches have specifically been explored to aid medical practitioners adhere to evidence-based medicine principles (13–15). These systems take queries (in natural language or key-terms) as input and generate/extract the query-relevant summaries. In terms of automatic summary quality, the performances of successful approaches designed for the medical domain have relied heavily on domain-specific knowledge sources (16). For example, the pioneering work by Demner-Fushman and Lin (17) incorporated sentence-level knowledge in a supervised classification system trained to detect *outcome* sentences, which were regarded as summary sentences. Sarker et al. (14) and ShafieiBavani et al. (15) utilized manually annotated summarization datasets to generate extractive and abstractive summaries—both systems relying heavily on the identification of domain-specific generalizations, concepts, and associations. Similarly, Hristovski et al. (18) proposed the use of domain-specific semantic relations for performing question answering for biomedical literature. Building on past research progress, recent studies have proposed end-to-end question-answering systems, which typically contain modules to perform the summarization (12, 19). Such systems, however, are generally only suitable for very specific types of queries, and despite their limited scopes, they invariably require the incorporation of medical domain-specific knowledge sources. The progress of summarization and question-answering research in the medical domain has been relatively sluggish, requiring considerable amounts of research efforts to overcome each of the many hurdles. Further discussion of the chronological progress in this research space is outside the scope of this brief research report, and detailed descriptions of medical domain-independent and domain-specific text summarization systems over the years are available through survey papers (20–22).

Adaptation of summarization systems to a particular domain can be computationally expensive and require large numbers of external tools (23). Within the medical domain, systems typically attempt to incorporate domain knowledge based on the Unified Medical Language System *via* software such as MetaMap (24), which can tag lexical representations of medical concepts. This is in turn used in downstream tasks, or as features in learning systems. Heavy dependence on these domain-specific systems introduces disadvantages, some of which are as follows:

(i) the systems are not very portable or generalizable, and are only suitable for the very specific tasks they were initially designed and evaluated for;
(ii) they are difficult to re-implement and/or deploy without the domain-specific knowledge sources or ontologies; and
(iii) they are computationally slow, often un-parallelizable.

The goal of our work is to design a resource-light and fast medical text summarization system that is decoupled from domain-specific knowledge sources. This work is an extension of our years of past research on this topic, focusing specifically on operational and deployment simplicity. The proposed system is extractive and query-focused in design. It relies on publicly available labeled data, which is used for weight optimization, unlabeled data—specifically, dense word embeddings learned from the unlabeled data—and a set of simple features that require little computational resources and time. In the development and evaluation processes, we selectively added and removed modules based on their performance and resource requirements. Comparative evaluation of our system against a state-of-the-art system on a standard dataset showed that it is capable of generating summaries of comparable qualities, despite its simplistic design.

METHODS

The primary dataset for this research is a corpus specifically for NLP research to support evidence-based medicine, created by Molla-Aliod et al. (25) with the involvement of the first author of this paper. The specialized corpus contains a total of 456 queries along with expert-authored single- and multi-document evidence-based summarized responses to them. Each query is generally associated with multiple single-document summaries, which present evidence from distinct studies. The abstracts of the studies from which the answers were derived are made available from PubMed. In total, the corpus contains 2,707 single-document summaries. To ensure fair comparisons, we used the exact train-test split from past research (14)−1,388 for

[1] https://www.ncbi.nlm.nih.gov/pubmed/ (accessed 25 Nov, 2019).

training and 1,319 for evaluation. The system we compare against is very reliant on domain-specific NLP resources, and it had produced state-of-the-art performance on the described corpus. The second dataset is much smaller, and we had prepared it to manually evaluate the performance of the summarization system. This dataset consists of a small set of articles describing research potentially relevant to COVID19. For each of these included articles, we manually created extractive summaries in response to a standard query, and we compared the agreement between our system and the manual summaries.

For developing and optimizing the system, we used the training set to devise feature scoring methods and learn weights for all the feature scores. Here, the training set does not consist of the exact single-document summaries, which are abstractive summaries authored by human experts. Instead, the training set consists of three-sentence extractive summaries for each document so that the gold standard is consistent with the expected output of our summarizer. These three-sentence extractive summaries of the training set are generated by computing the ROUGE-L F_1-score of all three-sentence combinations against the human summary, and selecting the top-scoring sentence combination for each text. We chose three as our target number of sentences based on past research (14, 17).

During the summary generation process, each sentence from the full set of candidate summary sentences receives a score for each feature included in the summarization system. All candidate sentences are then scored as the sum of the weighted feature scores, and the three sentences with the highest scores are extracted as the summary. The scoring process takes into consideration the target sentence position, the sentence length and the contents of the selected sentences. In the final summary, the selected sentences are presented sequentially (from first to last). The scoring process can be summarized as:

$$\zeta_{m,t_n} = \sum_{i=1}^{k} (w_{i,m,n} \times f_{i,m,n} | S_m, t_n)$$

where ζ_{m,t_n} is the score for sentence number m of a text, given the summary target sentence number t_n, and $w_{i,m,n}$ and $f_{i,m,n}$ are the weight and score for feature i, respectively. For each summary sentence position (t_n), the top-scoring sentence is chosen. To explore and discover a set of *simple* but *salient* features, we started with the full set of features used by the QSpec system and removed modules or features with the highest dependencies and longest running times. For example, one important derived feature in the QSpec system is a sentence-level score based on the *sentence type*, the *UMLS semantic types* present in the sentence, and the *associations between semantic types*. Identifying the sentence type requires the execution of an automatic classifier at run-time (26), identifying UMLS semantic types and associations requires the execution of MetaMap (24), and once these processes are completed, an exhaustive concept-level search is performed to find and score the sentence based on the presence of each association. Due to the computational complexity of this module, and its dependence on external tools, we removed this feature first and attempted to optimize performance using the other features—those attempted in the past and those we added. In addition to the features used for the QSpec system, we evaluated a number of features such as variants of edit-distance-based lexical similarities and scores based on the presence of possible statistical testing information (e.g., p-values). These features did not contribute to meaningfully improve the overall system score, and they were also eventually excluded. Following experimentation with multiple feature combinations, we selected five that could be computed fast and proved to be useful when used in combination. We describe these features in the following paragraphs.

Word Embedding-Based Maximal Marginal Relevance

Maximal Marginal Relevance (MMR) (27) is a strategy that can be used to increase relevance and reduce redundancy, and variants of it have been popular for text summarization (28–31). In our approach, we compute two similarity measures—between sentences and the associated query, and between the sentences themselves. During score generation, sentences are rewarded for being similar to the query, while at the same time they are penalized for being similar to sentences that have already been selected to be included in the summary. The similarity values are combined linearly with suitable weights (λ):

$$MMR = \lambda \times SIM(S_m, Q) - (1 - \lambda) \times \max_{S_c \epsilon S_{sel}} (SIM(S_m, S_c))$$

where $SIM(S_m, Q)$ is the similarity score between a sentence and the query and $\max_{S_c \epsilon S_{sel}}(SIM(S_m, S_c))$ is the maximum similarity between the same sentence and the set of already-selected summary sentences. Choosing the best three-sentence summary is a combinatorial optimization problem, and MMR enables us to approach sentence selection in a sequential manner. Despite the widespread use of MMR for extractive summarization, two variants of this score that we use in this system, which rely on distributed representations of the words in the sentences and the queries, had not been proposed in the past, to the best of our knowledge. We obtained pre-trained embeddings that were generated from all PubMed and PubMed Central (PMC) Open Access texts (32) using the *word2vec* tool[2] (vector size = 200, window size = 5) and the *skip-gram* model (33). For the first variant, we compute the similarity between two text segments (i.e., sentence vs. query and sentence vs. sentence) as the *average* cosine similarity of all the terms. We compute this average by adding the cosine similarities of all the term combinations and dividing by the product of the lengths of the two texts. For the second variant, we use the word vectors in a text segment to compute its *centroid* in vector space. A single centroid is computed for the set of all words within the set of already-chosen sentences (S_{sel}). These centroids are then used to compute MMR.

Traditional MMR Score

For the traditional MMR score, the third variant used in the system, we first pre-processed the terms by lowercasing, stemming and removing stop words. We then computed the *tf* × *isf* for each word in a sentence and the query—where *tf* is

[2] https://code.google.com/p/word2vec/ (accessed 17 Nov, 2019).

the frequency of a term in a text segment and *isf* is the *inverse sentence frequency* of the term in all the texts (i.e., the inverse of how many sentences, including the query, contain the term). We then generated vectors for each sentence using the $tf \times isf$ values of the terms.

Sentence Length Score

Sentence length is a metric that may filter out uninformative, short sentences by assigning them a lower score, while rewarding sentences that are relatively longer in a document. In summarization tasks where the character lengths of the summaries are limited, longer sentences may also be penalized (34, 35). We attempted to assign penalties to very short sentences (e.g., 1—3-word sentences), which often represent section headers. At the same time, our goal was to assign higher scores to longer sentences—with decreasing gradients for very long sentences, such that this score does not play a significant role in choosing between those informative sentences.

Our experiments on the training set suggested that a *sinusoidal* function conveniently served this purpose. The average sentence length in the training data is ∼150 characters, so we considered 0 and 300 characters to be the lower and upper length limits, respectively, and mapped the lengths to the range $\left(\frac{-\pi}{2}, \frac{\pi}{2}\right)$. Following that, we applied a *sin* function to the mapped value to generate a length score between (−1, 1). **Figure 1** illustrates how a sin function enables us to reward/penalize sentences based on their lengths relative to the average sentence length. Both reward and penalty start to level off as length approaches 0 or 300.

Sentence Position Score

Our last score is based on sentence position and the target sentence number. Sentence position has been shown to be a

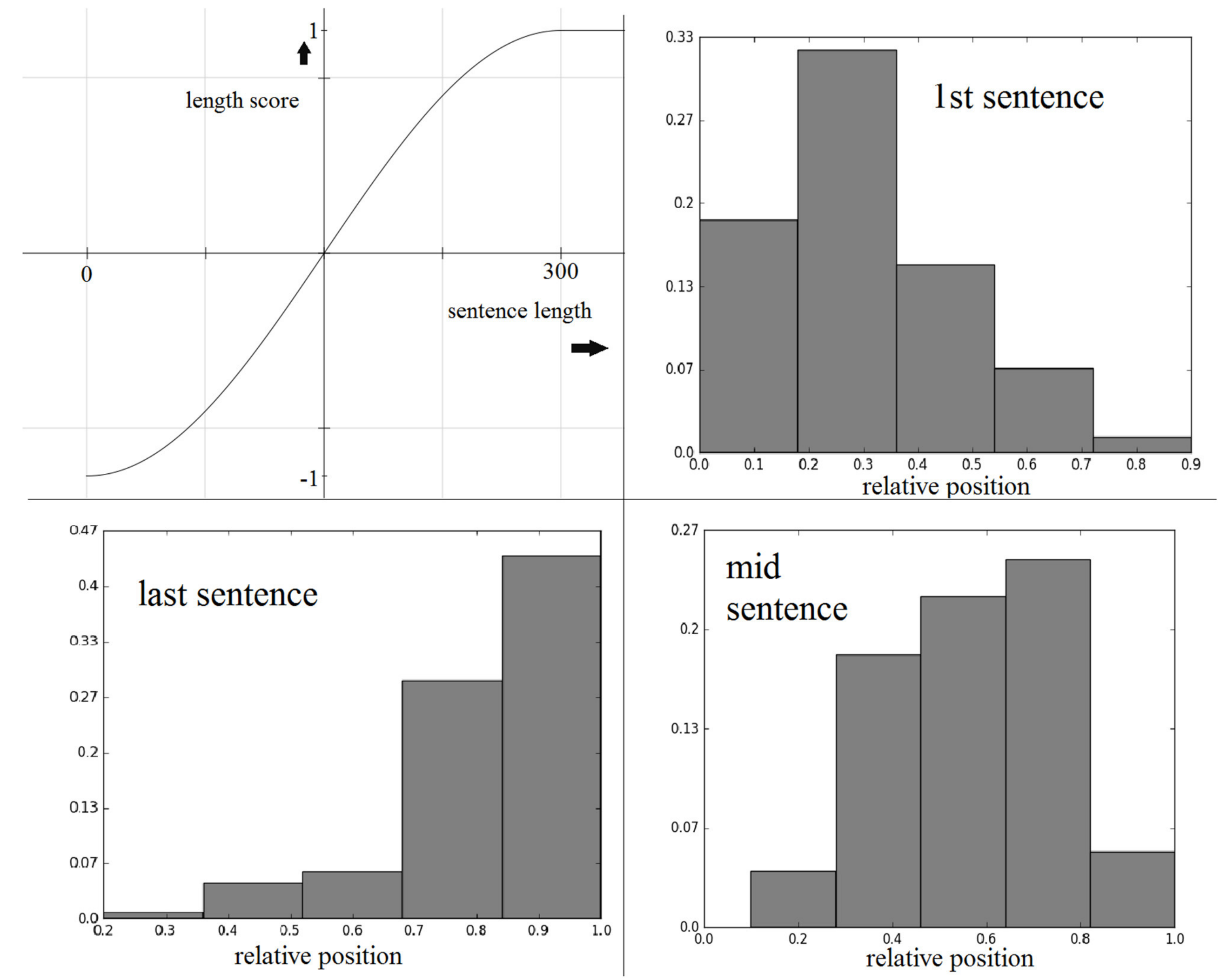

FIGURE 1 | Clockwise from top-left: sine function for sentence length score, maxed at 300 characters; first, middle, and last sentence relative position distributions from the best-scoring extractive training set summaries.

crucial metric for extractive summarization in domains including news (36) and medical (17). We used an approach identical to our past work as it had proven to be computationally fast and effective (14). The approach, which we called *target sentence specific summarization*, generates different scores for the same source sentence based on the summary sentence number. This means that the same sentence gets a different score when the system is searching for the first sentence for a three-sentence summary compared to when the system is searching for the last sentence. This ensures that the eventual summary extract is not biased to a specific region of the source text, which is often the case with traditional systems that apply the same scoring mechanism for all text spans. Generally speaking, when the system is scoring sentences for the first summary sentence, it gives preference to sentences occurring early in the source documents, which often contain important background information, compared to sentences occurring later, which tend to contain information about the final outcome of the study.

To compute this score, we first obtained the best three-sentence summary (gold standard summary) for each training text, and used these sentences to generate normalized frequency distributions of the relative sentence positions. These distributions are shown in **Figure 1**. During summary generation, given the relative sentence position *r* of a source sentence, the score assigned is the normalized frequency for *r* in the given target sentence distribution.

Weight Optimization and Intrinsic Evaluation

We computed near-optimal weights for scoring using the training set *via* a grid search in the range (0.0, 1.0). For each weight combination, all the three-sentence training set summaries were generated and the weights producing the highest F$_1$-score were used for evaluation on the test set. The ROUGE summary evaluation tool (37) was used to compare the extractive summaries with the expert-authored summaries in the corpus. The ROUGE-L variant of the evaluation tool attempts to score summaries based on their longest common subsequences (LCS) (38). Given two texts—the automatic summary of length *m* words and the corresponding gold-standard summary of length *n* words—the F$_1$-score is computed as:

$$F_1 - score = \frac{(1+\beta^2) \times R \times P}{(R + (\beta^2 \ \times P))}$$

where $R = LCS(summary, goldstadard)/m$; $P = LCS(summary, goldstandard)/n$ and $LCS(summary, goldstandard)$ is the length of the longest common subsequence between the summary and the gold standard. β^2 is set to 1. ROUGE scores had been shown to be correlated with human evaluators and the ROUGE-L F$_1$-score is the harmonic mean of the ROUGE-L recall and precision scores. In past research, the evaluations of many summarization systems were based on summaries constrained by character-level maximum lengths (e.g., 100 characters), and such evaluations typically used ROUGE recall scores for comparison. In our case, the summaries are constrained by the number of sentences (three), and so, optimizing and evaluating based on recall would overfit the system in favor of longer sentences. Therefore, we chose to use the F$_1$-score, rather than recall, and we computed them using the original human-authored summaries as the gold standard.

Extrinsic Evaluation on COVID19 Literature

We conducted a brief extrinsic evaluation of the system using a small number of recently-published articles about COVID19 or related research (39). We created six categories of queries focusing on different types of COVID19-related information (e.g., *treatment* and *transmission*). To establish these categories, we selected 2 from 12 categories that had been proposed in the literature (40) and added 4 additional ones. Two of our query categories (*treatment* and *prognosis*) overlapped with the categories proposed in the past, and we added 4 more categories based on their relevance to COVID19 and our research interests. Note that our intent was not to determine a comprehensive set of categories relevant to COVID19. The queries, their types and their numbers are shown in **Table 1**. For a set of 11 articles we manually created 3-sentence summaries. The four authors independently created the three-sentence summary for each article. We modeled the sentence selection task as a binary sentence labeling task and compared the pair-wise agreements between the annotators using Cohen's kappa (41).

We ran the summarizer with the best performing weights on a large amount of COVID19 literature that has been made available since the outbreak of the pandemic. For 11 of these articles, which were manually annotated, we compared the agreements between the three-sentence human summaries, and between the system and human summaries. We also compared the agreement between the human summaries and summaries generated by the QSpec system. Two authors were kept unaware of the internal scoring strategy of the system to ensure that sentence selection is not biased by that knowledge. Note that the articles themselves

TABLE 1 | Queries used for extrinsic evaluations, their types and numbers.

Query	Type	Number annotated
What measures can be taken to lower the transmission of COVID-19?	Transmission	3
What are some of the mental health impacts of COVID19?	Mental health	3
What is the common prognosis for COVID19 infection?	Prognosis	2
What treatment is effective for COVID19	Treatment	1
What are the common risks of COVID19 on targeted populations?	Risks	1
How does the impact of COVID19 vary based on social inequalities?	Inequalities/economic	1

were pre-selected, and were not based on the queries, since information retrieval is not an objective of our research.

EVALUATION AND RESULTS

Automatic Evaluations

Table 2 presents the performance of our system along with several other systems. Identical training-test splits were used for evaluation. Our proposed approach obtains a score of 0.166, 0.002 lower than the best-performing system. The table shows that despite the simplicity of our approach, its performance is comparable to the state-of-the-art, and significantly better than other baselines. To compare our approach with the extractive summarization method proposed by Demner-Fushman and Lin (17), we used an automatic classifier to detect *Outcome* sentences (42); the last three outcome sentences were extracted as the summary. Using the ROUGE score distribution of all summary combinations, we computed the percentile rank of our summarizer's performance *via* the method described by Ceylan et al. (43). In the proposed approach, a probability density function is generated using an exhaustive search of all ROUGE score combinations for extractive summaries, and this distribution is used to find the percentile rank of a system's ROUGE score. Our light-weight system's score has a percentile rank of 94.3 compared to QSpec's rank of 96.8. The large difference in percentile rank despite the small change in the ROUGE score is caused by the typical long-tailed nature of the ROUGE score distribution. The computed optimal weights for the features were: sentence position weight = 0.8; sentence length: 0.2; MMR (traditional): 0.2; MMR (dense vectors; both variants): 0.5.

Extrinsic Summary Evaluation

Pair-wise agreement, based on Cohen's kappa, was generally low for both sets of agreements (i.e., human-human and human-system). **Table 3** presents the average system-human, all human-human, and subsets of human-human agreements. Sample human and automatic summaries are provided in **Supplementary Material**. A link to the final human summaries, after resolving disagreements, are also provided in the **Supplementary Material**.

DISCUSSION AND CONCLUSION

Using a set of similarity-based and structural features, our system performs comparably to the state-of-the-art system, with a ROUGE-L F_1-score of 0.166. Our extrinsic evaluations showed that for this extractive summarization task, human-to-human inter-annotator agreement was low, resulting in a low ceiling for the automatic summarizer. We observed consistently low agreements across subsets of annotators, illustrating that choosing the optimal n-sentence query-focused summary is a difficult task for humans. Abstractive summaries could perhaps be more suitable for humans as more information can be summarized within a short text span. However, from the perspective of automatic summarization, moving from extractive to abstractive summarization has been challenging for this particular research community, and our scope was limited to extractive summarization. Although our evaluation was brief and differences between automatic and human summaries were not conclusive, we did observe more disagreements for earlier summary sentence selections compared to the selection of later sentences. Generally speaking, we found the gold standard summaries to have higher variance in relative sentence positions compared to automatically generated summaries. **Figure 2** further illustrates the differences between the gold standard extracts and the automatic summarization systems by visualizing the distributions of the relative positions of the sentences included in the summary. While human summaries almost invariably contain sentences from the end of the texts, they also tend to contain sentences from different relative positions. However, QSpec and our proposed system tend to select most sentences from the end and some from the beginning, but few from the rest of the document.

Our specific focus for this summarization system was to make the sentence selection process simple and decoupled from multiple additional systems or processes while also maintaining high performance. Our focus on simplicity is particularly from the perspective of deploy-ability (i.e., *how quickly can the source for the system be downloaded and executed on a new machine?*). Past systems focusing on the task of evidence-based medicine text summarization have relied on multiple knowledge-encapsulating software sources such as MetaMap, and parallel processes such as query and sentence classification, Compared to the resource-heavy QSpec system, which requires query and sentence classification and the generation of UMLS semantic types/associations, our current approach requires minimal pre-processing. Only a set of pre-trained word embeddings are required. The light-weight nature of the summarizer also means

TABLE 2 | Comparison of ROUGE-L F_1-scores for our summarizer with other systems and 95% confidence intervals.

System	ROUGE-L F_1 Score	95% CI
Our system	0.166	0.162–0.170
QSpec (14)	0.168	0.164–0.172
Last 3 Sentences	0.155	0.151–0.158
Demner-Fushman and Lin (17)	0.159	0.152–0.164
Random	0.154	0.150–0.157
First 3 Sentences	0.140	0.136–0.143

TABLE 3 | Average agreements between the system and human annotators, all human annotators and subsets.

Author/system	Kappa (average)
Human-system agreement average	0.33
Human-QSpec agreement average	0.35
Human-human agreement average	0.40
Average: annotators 1, 2 and 3	0.41
Average: annotators 1, 2 and 4	0.40
Average: annotators 1, 3 and 4	0.34
Average: annotators 2, 3 and 4	0.38

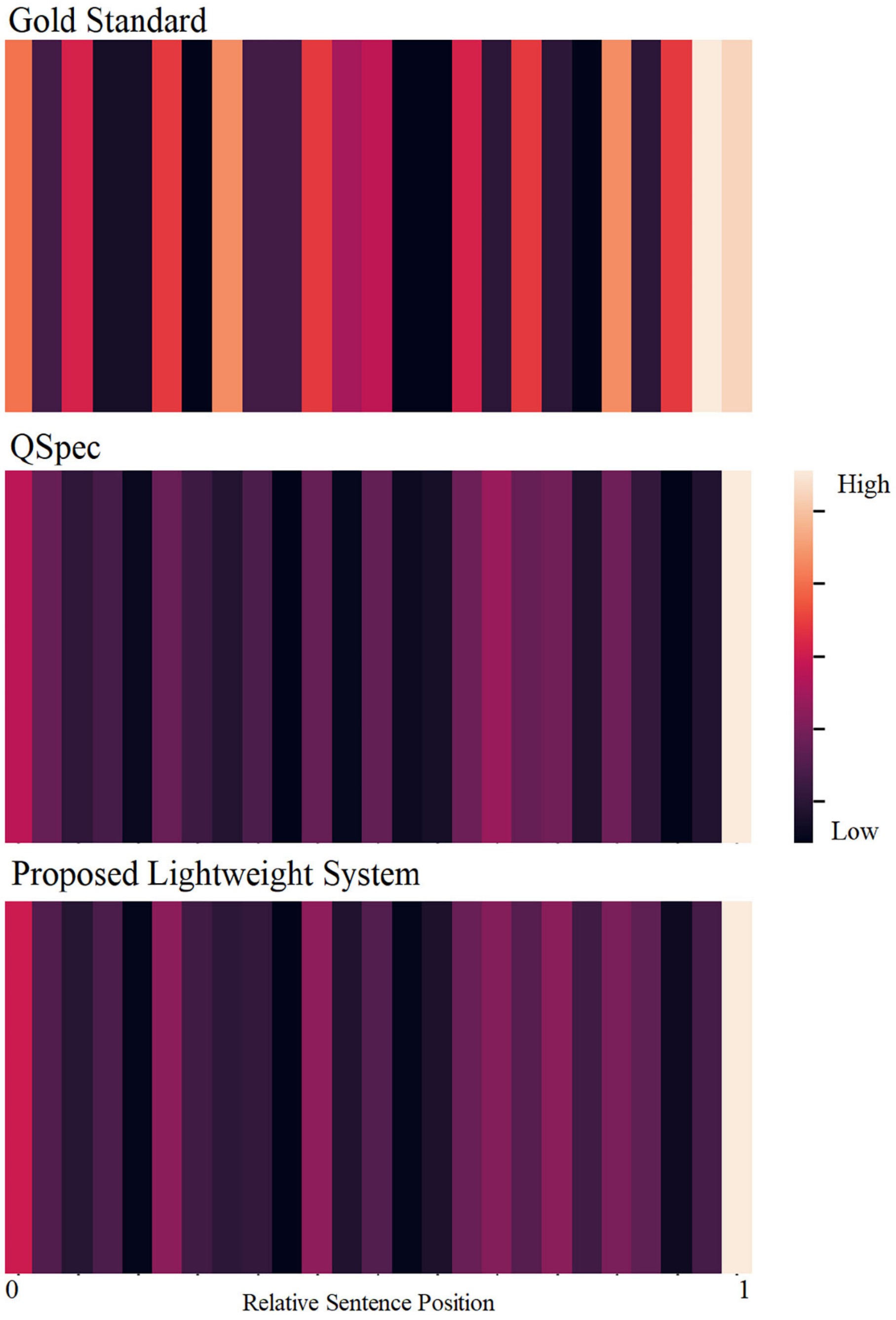

FIGURE 2 | Distributions of relative sentence positions of the gold standard summary sentences (top), and automatically-generated summaries (QSpec: middle; proposed system: bottom).

that it runs faster than QSpec. On a standard computer (Intel® i5 2.0 GHz processor), it takes our summarizer a few minutes to summarize all the documents in the test set. Due to the simplicity of our approach, we believe that it can be easily re-implemented, customized or extended for real-life settings, and the results can be reproduced without requiring the use of third-party tools. It is possible for non-NLP experts or even non-programmers to use the summarization system without having to set up additional tools; the only resource needed is any publicly available pre-trained word/phrase embedding model.

From an application perspective, we believe that this summarization approach is more easily *transferable* to other data sets, even those in other languages that do not have domain knowledge encoded in thesauruses. Exploring the applicability of this approach on non-English datasets is part of our future research plans. We are particularly interested in assessing the performance of this system, compared to those reliant on domain-specific knowledge sources, on other languages without including a language-specific gold standard or manually-curated knowledge sources. Our hypothesis is that this light-weight summarizer will outperform resource-heavy systems such as QSpec on such data sets.

We obtained the word embeddings from past research and used them without modification. There is a possibility that the learning of these embeddings can be customized to the summarization task for improving performance (e.g., using a COVID19-specific embedding model for the second summarization task). This is a notable limitation of the system—the semantics of emerging health topics, such as COVID19, may not be captured by the underlying embedding model, thus, leading to sub-optimal performance. Another operational limitation may be the size of the embedding model. While our focus is on a light-weight system that can be run on not-so-powerful computers, embedding models can be large in size (multiple gigabytes), which may act as an obstacle for old machines. To address these limitations, in future research, we plan to implement a continuously-learning embedding model that updates periodically using text from recently-published papers, and strategies for building targeted embedding models that require less unlabeled data and memory at run time.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://sourceforge.net/projects/ebmsumcorpus/ and https://sarkerlab.org/lw_summ/.

AUTHOR CONTRIBUTIONS

AS implemented the initial system and evaluated. AA, MA-G, and Y-CY assisted with the experiments, evaluations, and in preparing the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This research was conducted by generous funding provided by the School of Medicine, Emory University.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2020.585559/full#supplementary-material

REFERENCES

1. Djulbegovic B, Guyatt GH. Progress in evidence-based medicine: a quarter century on. *Lancet*. (2017) 390:415–23. doi: 10.1016/S0140-6736(16)31592-6
2. Grad R, Pluye P, Tang D, Shulha M, Slawson DC, Shaughnessy AF. Patient-oriented evidence that matters (POEMs)^{TM} suggest potential clinical topics for the Choosing Wisely^{TM} campaign. *J Am Board Fam Med*. (2015) 28:184–9. doi: 10.3122/jabfm.2015.02.140226
3. Sacristán JA. Patient-centered medicine and patient-oriented research: improving health outcomes for individual patients. *BMC Med Inform Decis Mak*. (2013) 13:6. doi: 10.1186/1472-6947-13-6
4. Greenhalgh T. *How to Read a Paper : The Basics of Evidence-Based Medicine*. 2nd ed. BMJ Books. London (2014). p. 222.
5. Greenhalgh T, Howick J, Maskrey N, Brassey J, Burch D, Burton M, et al. Evidence based medicine: a movement in crisis? *BMJ*. (2014) 348:g3725. doi: 10.1136/bmj.g3725
6. Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. 1996. *Clin Orthop Relat Res*. (2007) 455:3–5. doi: 10.1136/bmj.312.7023.71
7. Ely JW, Osheroff JA, Ebell MH, Bergus GR, Levy BT, Chambliss ML, et al. Analysis of questions asked by family doctors regarding patient care. *Br Med J*. (1999) 319:358–61. doi: 10.1136/bmj.319.7206.358
8. Ho GJ, Liew SM, Ng CJ, Hisham Shunmugam R, Glasziou P. Development of a search strategy for an evidence based retrieval service. *PLoS ONE*. (2016) 11:e0167170. doi: 10.1371/journal.pone.0167170
9. Methley AM, Campbell S, Chew-Graham C, McNally R, Cheraghi-Sohi S. PICO, PICOS and SPIDER: a comparison study of specificity and sensitivity in three search tools for qualitative systematic reviews. *BMC Health Serv Res*. (2014) 14:579. doi: 10.1186/s12913-014-0579-0
10. Swennen MHJ, van der Heijden GJMG, Boeije HR, van Rheenen N, Verheul FJM, van der Graaf Y, et al. Doctors' perceptions and use of evidence-based medicine: a systematic review and thematic synthesis of qualitative studies. *Acad Med J Assoc Am Med Coll*. (2013) 88:1384–96. doi: 10.1097/ACM.0b013e31829ed3cc
11. Hersh WR, Katherine Crabtree M, Hickam DH, Sacherek L, Friedman CP, Tidmarsh P. Factors associated with success in searching MEDLINE and applying evidence to answer clinical questions. *J Am Med Inform Assoc*. (2002) 9:283–93. doi: 10.1197/jamia.M0996
12. Sarrouti M, Ouatik El Alaoui S. SemBioNLQA: a semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions. *Artif Intell Med*. (2020) 102:101767. doi: 10.1016/j.artmed.2019.101767
13. Cao YG, Liu F, Simpson P, Antieau L, Bennett A, Cimino JJ, et al. AskHERMES: an online question answering system for complex clinical questions. *J Biomed Inform*. (2011) 44:277–88. doi: 10.1016/j.jbi.2011.01.004
14. Sarker A, Mollá D, Paris C. Query-oriented evidence extraction to support evidence-based medicine practice. *J Biomed Inform*. (2016) 59:169–84. doi: 10.1016/j.jbi.2015.11.010
15. ShafieiBavani E, Ebrahimi M, Wong R. Appraising UMLS coverage for summarizing medical evidence. In: *International Conference on Computational Linguistics (COLING)*. Osaka (2016). p. 513–24. Available online at: http://aclweb.org/anthology/C16-1050
16. Plaza L. Comparing different knowledge sources for the automatic summarization of biomedical literature. *J Biomed Inform*. (2014) 52:319–28. doi: 10.1016/j.jbi.2014.07.014
17. Demner-Fushman D, Lin J. Answering clinical questions with knowledge-based and statistical techniques. *Comput Linguist*. (2007) 33:63–103. doi: 10.1162/coli.2007.33.1.63
18. Hristovski D, Dinevski D, Kastrin A, Rindflesch TC. Biomedical question answering using semantic relations. *BMC Bioinfor*. (2015) 16:6. doi: 10.1186/s12859-014-0365-3
19. Yoo H, Chung K. PHR based diabetes index service model using life behavior analysis. *Wireless Pers Commun*. (2017) 93:161–74. doi: 10.1007/s11277-016-3715-9

20. Athenikos SJ, Han H. Biomedical question answering: a survey. *Comput Meth Programs Biomed*. (2010) 99:1–24. doi: 10.1016/j.cmpb.2009.10.003
21. Mishra R, Bian J, Fiszman M, Weir CR, Jonnalagadda S, Mostafa J, et al. Text summarization in the biomedical domain: a systematic review of recent research. *J Biomed Inform*. (2014) 52:457–67. doi: 10.1016/j.jbi.2014.06.009
22. Widyassari AP, Noersasongko E, Syukur A, Affandy Fanani AZ, Basuki RS. Literature review of automatic text summarization: research trend, dataset and method. In: *International Conference on Information and Communications Technology (ICOIACT)*. Yogyakarta (2019). p. 491–6. doi: 10.1109/ICOIACT46704.2019.8938454
23. Severyn A, Moschittiy A. Learning to rank short text pairs with convolutional deep neural networks. In: *SIGIR 2015—Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Santiago (2015). p. 373–82.
24. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc*. (2010) 17:229–36. doi: 10.1136/jamia.2009.002733
25. Mollá D, Santiago-Martínez ME, Sarker A, Paris C. A corpus for research in text processing for evidence based medicine. *Lang Resour Eval*. (2016) 50:705–27. doi: 10.1007/s10579-015-9327-2
26. Hassanzadeh H, Groza T, Hunter J. Identifying scientific artefacts in biomedical literature: the evidence based medicine use case. *J Biomed Inform*. (2014) 49:159–70. doi: 10.1016/j.jbi.2014.02.006
27. Carbonell J, Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval—SIGIR '98*. Melbourne, VIC (1998). p. 335–6.
28. Chandu KR, Naik A, Chandrasekar A, Yang Z, Gupta N, Nyberg E. Tackling biomedical text summarization: OAQA at BioASQ 5B. In: *BioNLP*. Vancouver, BC (2017). p. 58–66.
29. Mahajani A, Pandya V, Maria I, Sharma D. Ranking-based sentence retrieval for text summarization. *Adv Intell Syst Comput*. (2019) 851:465–74. doi: 10.1007/978-981-13-2414-7_43
30. Moradi M, Dashti M, Samwald M. Summarization of biomedical articles using domain-specific word embeddings and graph ranking. *J Biomed Inform*. (2020) 107:103452. doi: 10.1016/j.jbi.2020.103452
31. Wang B, Liu B, Sun C, Wang X, Li B. Adaptive maximum marginal relevance based multi-email summarization. In: Deng H, Wang L, Wang FL, Lei J, editors. *Artificial Intelligence and Computational Intelligence. AICI 2009. Lecture Notes in Computer Science, vol. 5855*. Berlin; Heidelberg: Springer (2009). p. 417–24. doi: 10.1007/978-3-642-05253-8_46
32. Pyysalo S, Ginter F, Moen H, Salakoski T, Ananiadou S. *Distributional Semantics Resources for Biomedical Text Processing*. (2012). Available online at: http://bio.nlplab.org/pdf/pyysalo13literature.pdf (accessed November 11, 2020)
33. Mikolov T, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: *Advances in Neural Information Processing Systems, Vol 2*. Lake Tahoe, NV (2013). p. 3111–9.
34. Abuobieda A, Salim N, Albaham AT, Osman AH, Kumar YJ. Text summarization features selection method using pseudo genetic-based model. In: *Proceedings—2012 International Conference on Information Retrieval and Knowledge Management, CAMP'12*. Kuala Lumpur (2012). p. 193–7. doi: 10.1109/InfRKM.2012.6204980
35. Ferreira R, De Souza Cabral L, Lins RD, Pereira e Silva G, Freitas F, Cavalcanti GDC, et al. Assessing sentence scoring techniques for extractive text summarization. *Expert Syst Appl*. (2013) 40:5755–64. doi: 10.1016/j.eswa.2013.04.023
36. Barzilay R, Mckeown KR. *Sentence Fusion for Multidocument News Summarization* (2005).
37. Lin C-Y. *ROUGE: A Package for Automatic Evaluation of Summaries* (2004).
38. Cormen TH, Leiserson CE, Rivest RL, Stein C. *Introduction to Algorithms*, 3rd ed. Cambridge, MA: The MIT Press (1989).
39. The Whitehouse. *Office of Science and Technology Policy. Call to Action to the Tech Community on New Machine Readable COVID-19 Dataset*. The White House (2020). Available online at: https://www.whitehouse.gov/briefings-statements/call-action-tech-community-new-machine-readable-covid-19-dataset/ (accessed November 11, 2020).
40. Yu H, Cao YG. Automatically extracting information needs from Ad Hoc clinical questions. *AMIA Annu Symp Proc*. (2008) 2008:96–100.
41. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. (1960) 20:37–46. doi: 10.1177/001316446002000104
42. Kim S, Martinez D, Cavedon L, Yencken L. Automatic classification of sentences to support evidence based medicine. *BMC Bioinform*. (2011) 12(Suppl. 2):S5. doi: 10.1186/1471-2105-12-s2-s5
43. Ceylan H, Mihalcea R, Özertem UU, Lloret E, Palomar M. Human quantifying the limits and success of extractive summarization systems across domains. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*. Los Angeles, CA (2010). p. 903–11.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.