Automated Text Summarization of Research Papers Regarding the Effectiveness of Various Treatment Plans for Leukemia

Jan Apolline D. Estrella
Philippine Science High School
Agham Road, Diliman
Quezon City, 1101 Metro Manila
+639335607689, PH
b2019.apolline.estrella@pshs.edu.ph

Christian Philip L. Gelera
Philippine Science High School
Agham Road, Diliman
Quezon City, 1101 Metro Manila
+639470447218, PH
b2019.tayan.gelera@pshs.edu.ph

Christian S. Quinzon
Philippine Science High School
Agham Road, Diliman
Quezon City, 1101 Metro Manila
+639369951454, PH
b2019.christian.quinzon@pshs.edu.ph

Ethel C. Ong
De La Salle University
2401 Taft Avenue, Malate
Manila, 1004 Metro Manila
(02) 524-0402, PH
ethel.ong@delasalle.ph

Mc Jervis S. Villaruel
Philippine Science High School
Agham Road, Diliman
Quezon City, 1101 Metro Manila
+639778220413, PH
mjsvillaruel@pshs.edu.ph

Edlen Mari M. Sanchez
Philippine Science High School
Agham Road, Diliman
Quezon City, 1101 Metro Manila
+639178287458, PH
emmsanchez@pshs.edu.ph

ABSTRACT

Text summarization involves the identification and extraction of sentences from long documents to produce shorter-length summaries that enable readers to have quick access to relevant information. In this paper, we describe our automated text summarizer that is capable of extracting essential information from research papers regarding various treatment plans for leukemia to aid users in reviewing online research publications more efficiently. The summarizer applies natural language processing and sentence scoring techniques to select parts of a paper, e.g. introduction, body, results, in bullet form. The generated summaries from three leukemia-related papers were then subjected for manual evaluation by experts and laymen. These summaries were found to consist of concise, informative and self-contained descriptions regarding the abstract and the introduction of a given article. However, these were also observed to have an excess of presented data, and a lack of background information, which an end-user is mostly likely required to know to fully comprehend the methods and findings of a study.

Keywords

Automated text summarization, sentence extraction, sentence scoring, cancer, leukemia, treatment plans

1. INTRODUCTION

One of the leading diseases in the world is cancer [40], which is a collection of related illnesses that are characterized by out-of-control cell growth. In a study of Ferlay et al. [2014], there were an estimated 14.1 million new cancer cases and 8.2 million cancer deaths that occurred in 2012 worldwide alone.

Most cancers cannot be cured as most are chronic diseases [28]; one of the aforementioned illnesses is leukemia, the most common cancer in children, accounting for about 30% of all pediatric cancers [7]. Fortunately, chronic cancers may be controlled for periods of time with treatment. These treatment plans vary in terms of effectiveness, depending on how cancer would respond to the given medication [28].

For a person with cancer, making treatment choices is an extremely crucial process as it would determine how one would survive. For almost each type of cancer, there are various treatment plans to look into. Choosing which treatment one will undergo involves several factors such as the type of disease, location of the cancer, amount of cancer, extent of spread, one's overall health and personal wishes [28]. More importantly, the efficacy of a treatment plan will greatly influence one's decision. Aside from choosing a treatment, it is also important to make treatment choices as fast as possible before cancer has the chance to spread, making treatment become more difficult, thus possibly reducing a person's chances of surviving [6].

Treatment choices are best discussed with one's doctor, cancer care team and family members. Some also choose to join face-to-face or online support groups in which they are able to meet people with varying experiences and backgrounds in relation to decisions regarding treatment plans [3]. Through sharing knowledge regarding the effectiveness of their treatments, people can help one another in their own treatment choices. To gather more information, several choose to review different written materials, such as health information packages, leaflets and research articles to study and compare the efficiency of different treatments.

It was also found by Coulter and Ellins [2007] that motivating patients to be more participative and active in decision-making regarding their health has great effect on quality, efficiency, and results, thus strategies should be formulated in order to stimulate and enhance effective patient engagement. An initiative in the United States that gave computer-based support for underprivileged masses worked most effectively as compared with other approaches, as there was more health information for them to access. Therefore, it was concluded that, to encourage patients in personally contributing to their own treatment choices, health literacy must be continually enhanced and developed by means of providing informative decision aids, such as the aforementioned pamphlets, and computer-based and online health information, which include research articles and research journals. More importantly, in the study of Reeve, Han and Brooks [2007],

Philippine Computing Journal Dedicated Issue on Natural Language Processing, pages 21–28
Vol. XIII No. 1 August 2018

it was stated that physicians, such as oncologists, must constantly study clinical trial studies that are related to their field of specialization in order to improve patient treatment.

Unfortunately, the quantity of biomedical information found on the Web continues to increase exponentially, making the process of studying multiple research studies in order to review different kinds of treatment time-consuming [34]. Specifically, one can find over 13,500 clinical trials in the US National Institutes of Health Clinical Trials database, and over 16 million citations from 4800 journals in PUBMED [24]. However, with the use of automated text summarization, the quantity of text found in a lengthy research paper can be reduced while preserving its core information [21].

Automated text summarization is the process of using natural language processing techniques in the identification of the most important data from a source to produce a concise representation of the text input [21]. It has been utilized mostly in processing huge volume of news articles, but over the years, has also shown great potential in the extraction of relevant information from medical literature: text summarization is able to aid those in the area of biomedicine who have to read through various research articles and journals. Reduction of data found in biomedical papers has been found to increase productivity by aiding experts in efficiently finding pertinent texts, and by presenting only the important information from these documents with decreased effort [26]. Several significant works that showed the remarkable capabilities of automatic text summarizers in the medical field include that of Afantenos, Karkaletsis, and Stamatopoulos [2004], who created a system that can produce a single summary from multiple documents in the pharmaceutical industry, and Aramaki et al. [2009], who programmed a software that can convert a medical text into a tabular structure. Though the usage of automated text summarization is prevalent in the biomedical domain, there is a lack of studies that specifically concern the effectiveness of extractive automatic text summarizers in the reduction of data found in research studies that contain cancer- or leukemia-specific concepts.

In this paper, we describe our approach in building an automated text summarizer that can process single documents to generate extractive summaries by means of concepts that are domain-specific to leukemia.

2. RELATED LITERATURE

2.1 Leukemia and its Treatment

Leukemia is the most common cancer in children, accounting for about 30% of all cancers in children [27]. Several types of leukemia, such as chronic myeloid leukemia and chronic lymphocytic leukemia, are chronic illnesses [28]. Despite being incurable, these chronic diseases can be controlled or managed over long periods of time by means of treatment.

Making treatment choices determines one's chances of survival; therefore, choosing which treatment one will undergo involves several factors such as the type of disease, location of the cancer, amount of cancer, extent of spread, one's overall health and personal wishes, and most importantly, the efficacy of a treatment plan [27]. These decisions are seen to be best discussed with one's doctor, cancer care team and family [28]. Some may even choose to join face-to-face or online support groups in which one can meet people with varying experiences and backgrounds in relation to decisions regarding treatment plans [3]. To gather more information, several choose to look through various reading materials, such as health information packages and leaflets to study and compare the efficiency of different treatments.

2.2 Automated Text Summarization

Automated text summarization is the extraction of relevant data from a document to produce a concise representation of the text input. In the recent decades, this process has been further used and developed to reduce the quantity of text found in a lengthy research paper while preserving its core information [34].

2.2.1 Characteristics

Automated text summarizers have different distinguishable features. The existence of such characteristics can be owed to the varying inputs and outputs one can respectively feed to and obtain from an automated text summarization software [11].

Text summarization algorithms vary due to the input that can be received: the input of a text summarizer can either be a single document, which is considered as a single input text, or a multi-document, which is considered as one text that covers the content of multiple input texts. Additionally, the input of a text summarizer can either be domain-specific, focusing on specific content, or general, accepting input text from any domain [11].

Additionally, text summarization softwares differ according to the output generated. The output of a text summarizer can either be an extract composed of a collection of passaged extracted from the input text, or an abstract, which is an output text newly generated by the program. The output of a text summarizer can also either be a fluent and coherent summary that consists of full, grammatical sentences, or a disfluent and fragmented summary that consists of individual words or portions of text that do not constitute grammatical and coherent sentences or paragraphs [11].

2.2.2 Applications in Various Fields of Specialization

Automated text summarization techniques are found to be useful in various fields of specialization. Businesses and corporations utilize extractive text summarization, accompanied by speech recognition, for the productive generation of "meeting minutes". On the other hand, legal experts use abstractive text summarizers for the efficient compression and restatement of content found in important judicial papers regarding laws, court decisions, and tribunal processes. Extractive and abstractive text summarization techniques can also be concurrently implemented: online news portals, such as Google and Columbia Newsblaster, use both the aforementioned methods in order to provide users a brief summary of a news article [35].

In the field of biomedicine, effective automated text summarizers have been developed to aid experts in studying a massive volume of data to improve healthcare and efficiently make treatment choices. Such include *Text Retrieval Extraction and Summarization Technologies for Large Enterprises* (TRESTLE), a system that can produce a single-sentence summary of documents used in the pharmaceutical industry by means of Information Extraction (IE) technologies involving the usage of domain-specific features [34]; *AskHERMES*, a question-answering system that can answer complex clinical questions with extractive summaries [43]; *Text2Table*, a software that can summarize by converting a medical text into a tabular structure using extraction strategies [12]; and *BioChain*, *FreqDist*, and *ChainFreq*, novel extractive text summarization algorithms that were developed to identify significant sentences in biomedical research articles [24].

For the development of all kinds of automated text summarizers, natural language processing techniques are employed for the

effective and efficient processing of data found in documents, one of which is text [36].

2.2.3 Qualitative Evaluation of Automated Text Summarizers

As stated by Lloret and Palomar [2009], the qualitative evaluation of automated text summarizers is important. The primary underlying purpose of qualitative evaluation is to establish sets of criteria pertaining to the quality of the summaries without the usage of reference models. The aforestated criteria should include summary coherence, topic identification, coreference resolution, summary informativeness, and natural language generation. In the study of Mani and Maybury [2001], it was also found that assessors can qualitatively evaluate summarizers by using a Likert scale to rate software-generated summaries according to criteria that pertain to their preferred length, intelligibility, and perceived usefulness. Additionally, according to Saziyabegum and Sajja [2017], qualitative evaluation of automated summarizers must consider the following factors that are used in Document Understanding Conferences and Text Analysis Conferences: redundancy, grammaticality, referential clarity, and structure and coherence.

2.3 Natural Language Processing

Natural language processing (NLP) is the field of study in the area of computer science that focuses on the emergence and development of technologies that can understand human natural language with the use of computational linguistics and artificial intelligence [15].

Several tools have been developed by various researchers for the creation, publication, and manipulation of linguistic data over the past decades. Written in Python, the Natural Language Toolkit (NLTK) is used for working with natural or human language with interfaces such as WordNet, an English lexical database, and other text-processing instruments for classification, tokenization, stemming, tagging, parsing, and semantic reasoning [37]. In addition, it is a library containing program modules, datasets, and tutorials on NLP.

Using these NLP tools, there are various NLP techniques that can be utilized in the implementation of automated text summarization: some of which are tokenization, part-of-speech (POS) tagging, WordNet, and sentence scoring.

Tokenization is a process or tool in Natural Language Processing that breaks a block of text into tokens, which can represent symbols, letters, words, phrases, or sentences. Moreover, POS tagging can be executed using the POS tagger, a tool that parses input human language text and outputs an annotated version of it based on human grammar. Furthermore, WordNet is an online lexical database for the English language containing information about 155,000 different parts of speech and including simplex, phrasal verbs, and idioms, useful for text disambiguation and classification, and information retrieval. Lastly, sentence scoring is a method of identifying the most relevant sentences in a text document [33].

2.4 Sentence Scoring

Sentence scoring techniques, which are implemented for the development of extractive text summarizers, are utilized in the determination of significant sentences in a text. The scoring of sentences can be performed using three main approaches: word scoring, which designates scores to the most relevant words; sentence scoring, which assigns scores to sentences based on their features; and graph scoring, which analyzes the interrelations between different sentences [33].

Different scoring techniques utilize various sentence features. Some, such as term frequency-inverse document frequency, consider the frequencies of each word within a sentence and within the whole document. Sentence position considers the position of the sentences in regards to its parent paragraph. Sentences containing relatively high quantities of numerical data may also contain more pertinent data towards the summary, and are given a corresponding score. Similarly, cue-phrase scoring considers the number of keywords a sentence contains based on a list of keywords determined previously; the higher the number of keywords, the higher the score. The sentence's similarity to the title may also be utilized on account of the title containing relevant keywords vital to the topic of the journal. Aggregate similarity scoring uses similarities between pairs of sentences to generate weights for each sentence [36].

No single scoring technique is able to accurately determine the relevance of a sentence. Moreover, since different scoring techniques work on various sentence features, a system of varying sentence scoring techniques is commonly implemented.

3. METHODOLOGY

For the analysis of leukemia-related research studies prior to the development of the algorithm, a corpus comprising of 30 research articles in portable document format (PDF), regarding treatment plans for leukemia were obtained from credible medical databases such as ProQuest, EMBASE, MEDLINE, and the National Center for Biotechnology Information (NCBI) [9] by using "*treatment for leukemia*" as the search key. Through examination, each article was found to have a common document structure of significant segments; abstract, introduction, methodology, and findings. The articles were then thoroughly analyzed to determine significant cancer-related keywords or key phrases. Additionally, the taxonomy defined in the National Cancer Institute (NCI) Dictionary of Cancer was used to define supplementary key terms that are specific to the domain of oncology, such as the following: *survival rate*, *overall survival rate*, *five-year survival rate*, *disease-free survival rate*, *complete remission*, *complete response*, *partial remission*, *partial response*, *rate of relapse*, and r*elapse rate*. These keywords and key phrases were later used in one of the sentence scoring techniques, namely cue-phrase scoring technique.

The algorithm is programmed in Python using the Python Integrated Development and Learning Environment, along with the Natural Language Toolkit (NLTK) module. As shown in Figure 1, the process of developing an automated text summarizer consists of three phases; preprocessing, sentence scoring, and summary generation.

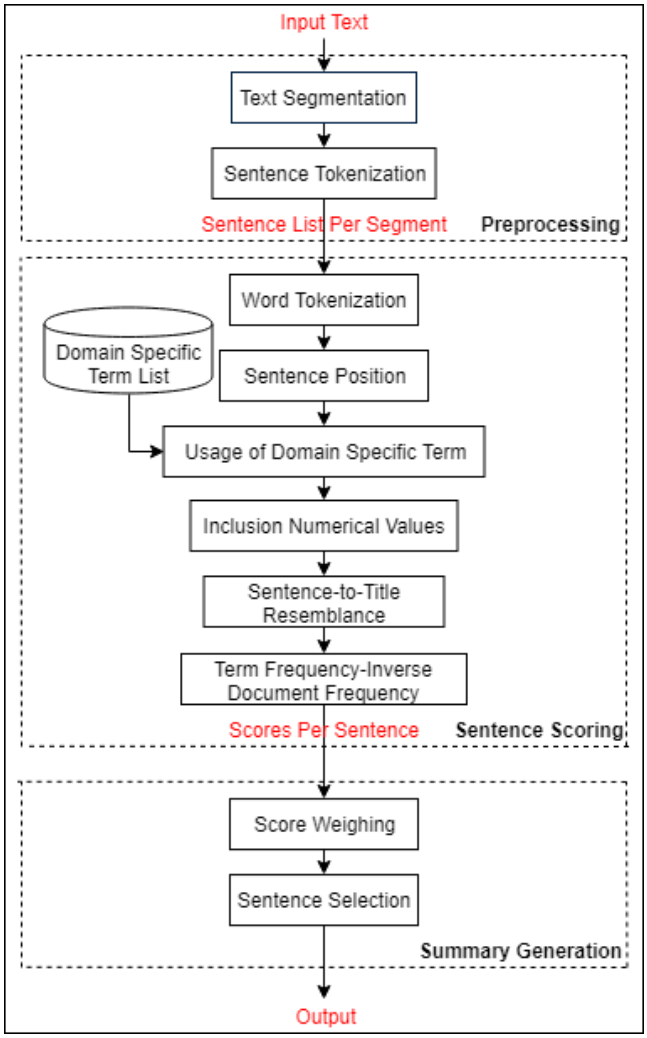

Figure 1. Methodology

3.1 Preprocessing

Certain text extraction hindrances, such as the inclusion of headers and footers, and errors caused by document layouts that affected the order of sentences, resulted from directly extracting text from PDF files. This required the text pertaining to the article of each document from the corpus of procured articles to be manually transferred to respective unicode-formatted text documents.

The first phase of the automated text summarizer program is the segmentation or the division of a research paper into a list of its respective sections (e.g. abstract, introduction, methodology, findings). Using the NLTK module *tokenize* package's sentence and word tokenizer, the algorithm then separates the input document into a list of sentences, and a seperate list of words and symbols. Stop words, which are common words with little to no relevance to any given text, were eliminated. These were obtained from the list of stop words provided by one of the text corpora found in the NLTK.

3.2 Sentence Scoring

After text segmentation and tokenization, a list or array is initialized, each element of which corresponds to a sentence's score that determines the 'significance' of a sentence. To identify significant sentences in a given input article text, sentence scoring was performed using a variety of sentence-ranking methods, namely sentence position, cue-phrase sentence scoring, inclusion of numerical data, similarity to title, and term frequency-inverse document frequency (TF-IDF). Each technique provides a score to each sentence; cumulatively, these points determine the significance of each sentence.

Sentence position. This feature can be used to give sentences closer to the top and bottom of a segment a score, and give a score of zero to the rest of the sentences. The technique assumes that sentences near the top and bottom of a segment contain key ideas about that segment. In the algorithm sentences that are part of the top 20% and bottom 20% sentences of a segment, a score of 1 is given [33].

Cue-phrase sentence scoring. This technique assigns scores to a sentence based on how many keywords or key phrases the sentence contains. Cue phrases include keywords significant to the study and transitional devices [33]. A cue-phrase ontology of scientific terms that are related to treatment plans for leukemia was constructed from the analysis of the procured articles: such words include 'survival rate', 'complete remission', and 'progress rate'. The formula for cue phrase scoring is:

$$CP = \frac{CPS}{CPD}$$

where CP = cue-phrase score,
CPS = the number of cue phrases in a sentence, and
CPD = the total number of cue phrases found in a document.

Sentence inclusion of numerical data. This scoring assumes that sentences with numerical data are more relevant since they contain quantitative data that may be part of the results or could be significant to a process in the treatment [33]. Such data may refer to quantities relevant to the topic, such as *rate of remission*, *medicine dosage*, or the *survival rate of a treatment plan*. The formula for sentence inclusion of numerical data is:

$$Score = \frac{log\,(10n)}{log\,15}$$

where n = the number of quantitative data in a sentence.

Sentence similarity to title. This subscoring technique calculates for a sentence's similarity to the title [33]. This scoring assumes that words found in the title are relevant to the topic, and sentences more similar to the topic are consequently more relevant. The formula for sentence similarity is:

$$Score = \frac{Ntw}{T}$$

where Ntw = the number of the title's constituent words in the sentence, and
T = the number of the title's constituent words.

Term frequency-inverse document frequency (TF-IDF). For each word in a sentence, a score is given based on how often the word occurs within the sentence and within the document [33]. Specifically, TF-IDF calculates for the relevance of a word to the whole document. The overall weight of each sentence is the total score of all its constituent words. The formula for TF-IDF is:

$$TF - IDF = SN(\frac{log(1+tf)}{log(sf)})$$

where N = the number of sentences,

tf = the frequency of a term in a sentence, and
sf = the frequency of a term in all sentences.

Scores from this technique had a significantly higher range than those of the other techniques. This required the obtained scores from this equation to be normalized into an interval of zero to one with the equation:

$$F_n = \frac{TF - IDF_n}{M}$$

where F_n = the final TF - IDF score
TF - IDF $_n$ = the TF - IDF score of sentence n
M = Top TF - IDF score of the document

3.3 Summary Generation

Each sentence scoring technique is applied to every sentence, and the cumulative scores for each of the sentences are gathered. The sentences are ranked among each segment of the article. The highest-ranking N percent of each segment or section is included in the generated summary, where N is a number predetermined by the researchers. The value of N varies for each section of the article: 40 for the abstract, 35 for the introduction, and 15 to both the methodology and the findings. The percentages were determined based on the observed general number of sentences found in each each segment of an article. The segments with relatively fewer sentences are assumed to be more concise.

Listing 1. Sample structure of a generated summary.

I. Abstract
   - In the present study, we describe the successful treatment of a 71-year-old Japanese female patient with Ph?MPAL by the alternation of second-generation tyrosine kinase inhibitors according to BCR-ABL1 mutations.
   - The patient survived in her third complete remission (CR) for over 4 years.
   - …

II. Introduction
   - MPAL is designated as a disease entity in the revised version of World Health Organization's (WHO) Classification of Tumors of Hematopoietic and Lymphoid Tissues in 2008.
   - ….

III. Methodology
   - *Sentence 1*
   - *Sentence 2*
   - *Sentence 3*
   - *…*
   - *Sentence n*

IV. Findings
   - **...**

4. TEST RESULTS AND ANALYSIS

Qualitative evaluation of three summaries generated by the automated text summarizer were performed by two NLP experts and two laymen through manual assessment and comparison of the software-generated summaries vis-a-vis the original article. As adapted from the methods of Solis, Siy, Tabirao, and Ong [2009], a set of criteria was defined for use in the qualitative evaluation. Using the definition of a good summary from the International Labour Organization [2005], each output was evaluated along two main criteria -- content and language using a Likert scale of 1 - 5, (5 - strongly agree, 4 -- agree, 3 -- neutral, 2 -- disagree, 1 -- strongly disagree).

The *content* criterion looks at the following characteristics of a good summary [37]: it must include only the core information; it must have a structure (e.g. introduction-body-conclusion); it must present the rationale or purpose, results, conclusions, and/or recommendations of a paper. Additionally, the summary should be able to state the main topic or idea of the paper. On the other hand, *language* criterion focuses on the following aspects of the summary: its constituent words should be easy to understand by end-users who can have no complete background knowledge on the main idea of the paper; its sentences are clear and have simple structures; and its presented results are accurate.

4.1 Evaluation of the Content

The mean scores obtained from the evaluation of the content of all the summaries are shown in Table 1. The performance of the automated text summarizer in terms of its contextual aspect was given a general average score of 4.08.

The capability of the summarizer to produce summaries that follow a consistent structure received a perfect score of 5.00. A rating of 4.64 was received by the software for its accuracy. With relatively high scores of 4.47, 4.28, and 4.14 respectively, the summaries were also found to be able to present a concise description that would represent the abstract of a study, to state the main idea of an article, and to summarize the introduction to a research. Most importantly, it was agreed by the evaluators that the summaries clearly state the benefits and potential risks of a treatment plan as these respondents gave an average score of 4.06 for the corresponding criterion. Furthermore, the automated text summarizer obtained 3.61, 3.47, and 3.08 for its capacity to provide information regarding the methods for conducting a study, to discuss the results obtained by a research, and to provide only the important information.

Table 1. Mean Score Given to all the Summaries for the Evaluation of their Content

Criteria	Average
Summary only includes important information.	3.08
Summary follows a consistent structure: <abstract-introduction-methods-findings>.	5.00
The Abstract gives a concise description of the research article.	4.47
The Introduction states the background information of the research, the purpose of the study, and a brief overview of the methodology.	4.14
The Methods section provides information about how the study was conducted.	3.61
The Findings section presents and discusses the results obtained.	3.47
Summary contains accurate information.	4.64
Summary clearly states the main idea of the article.	4.28
Summary clearly states benefits and potential risks of the treatment plan.	4.06
General Average	**4.08**

The generated summaries were given relatively higher ratings in terms of their structure. This is likely attributed to the presence of the required sections (e.g. abstract, introduction, methods, findings) in all the articles that were part of the test corpus.

Evaluators also observed that the summaries contain accurate information. This may be attributed to the direct extraction of relevant content rather than abstraction of text from the source articles, thus presenting correct information.

The capability of the summarizer to create a concise description that would represent the abstract of a study, to state the main idea of an article, and to summarize the introduction to the study received high ratings, relative to that of the others. These scores can be owed to the algorithm that utilized the resemblance of a sentence to the title of a journal publication as the most important sentences are most likely related to the title of the text they belong to [42].

The evaluation shows not only that the summarizer built specifically for research articles on leukemia treatment plans was competent, it was also observed that the summaries showed the benefits of a treatment plan well. However, several evaluators noted the failure of the software to present potential risks of a treatment. This may be due to its lack of scientific terms that belong to its ontology of cancer-related terms, which is being used for the scoring of sentences based on cue-phrases. These words should not only be specific across a specific domain, which, for this study, should be associated with treatment plans for leukemia, but also across the main topic that a research study covers as the cue-phrase sentence-scoring feature heavily relies on the prior preparation of a set of keywords or key phrases [42].

The scores from the evaluation of the generated summaries' methods and findings section were also relatively lower than of the other sections. This may be due to the lack of definitions or details of cancer-specific concepts and terms needed for oncology laymen to completely understand the treatment. As aforestated, the software extracts only the top n sentences per section of the paper, therefore possibly not being able to present all important information that is needed to understand the processes used in a research study. Several terminologies that were included in the summaries provided regarding the findings of a study were also found to be difficult to understand by people who are non-experts in the field of biomedicine or leukemia. In addition, these ratings may have been caused by the inaccuracy of term frequency–inverse document frequency approach as there is usually a massive volume of stop words, or terms with little to no relevance to the study, and the list of stop words that was obtained from the NLTK, which was used in the aforementioned technique, may have lacked stop words that are specific to the domain of leukemia treatment plans [33].

Lastly, the effectiveness of the summarizer in including the important information only was given an average score that is low, as compared to the other ratings. The evaluators observed that the algorithm failed to omit excess and irrelevant data. The excess of data and lack of background information may have originated from the existence of synonyms, which may greatly affect the score of a word given by different sentence-ranking techniques, such as those that involve the use of cue-phrases and term frequency.

In general, the evaluators, who provided a general mean score of 4.08, were to found to agree that the software is an effective automated text summarizer for research papers regarding treatment plans for leukemia in terms of its contextual aspect.

4.2 Evaluation of the Language

The mean scores obtained from the evaluation of the linguistic aspect of all the summaries are shown in Table 2. With a general average score of 3.36, the summaries were given relatively lower ratings in terms of their linguistic aspect, as compared to that of their content.

Specifically, the software garnered a rating of 3.58 for both its understandability and ability to create summaries that do not assume complete background knowledge from an end-user. Additionally, a rating of 2.92 was given by the evaluators for the presentation of sentences by the summaries.

Table 2. Mean Score Given to all the Summaries for the Evaluation of their Linguistic Aspect

Criteria	Average
Summary is understandable.	3.58
Summary, as a whole, does not assume complete background knowledge from the reader.	3.58
Each sentence is neatly presented.	2.92
General Average	**3.36**

The score obtained by the summarizer for the understandability of the summaries received a score of 3.58; this may be due to the occurrences of sentences whose context rely on other statements that were not extracted by the summarizer [29]. It was also found that the summarizer assumes complete background knowledge from the reader, as sentences were only extracted from a source publication, and there was no simplification of words and sentences performed. It was suggested by some evaluators that the simplification of information presented may be beyond the scope of the current research project as its current objective is to only extract significant sentences that contain that most essential information.

Finally, the presentation of sentences was given a lower score, as compared to the previous ratings. It was observed that some words were split into two parts due to being at the end of a line in the original copy of the article (e.g. admin- istration, mechanisti-cally). Citation indices (e.g. [7-9]) and in-text citations were also included, potentially hindering people who are not from the academe from fully understanding a summary. Several sentences were also observed to have merged together. A number of symbols were seen to be misplaced and not properly converted.

In essence, the evaluators, who delivered a general mean score of 3.36, were to found to be neutral towards the effectiveness of the software as an effective automated text summarizer for research papers regarding treatment plans for leukemia in terms of its linguistic aspect.

5. CONCLUSION AND FURTHER WORK

In recent years, automated text summarization by means of natural language processing has increasingly gained attention. By condensing the data contained in clinical studies, end-users, such as laymen and physicians, have found it easier to study a massive volume of information [26].

Using sentence-scoring techniques, a working algorithm for the automated text summarization of research articles regarding the effectiveness of various treatment plans was implemented. Results obtained from the qualitative evaluation of the software show that the automated text summarizer for research papers regarding the effectiveness of various treatment plans was successfully

developed: the software was effective in presenting the benefits and potential risks of a treatment plan and generating summaries from research papers regarding treatment plans for leukemia in terms of its contextual aspect.

Due to the low ratings received by the software in terms of its linguistic aspect, researchers who have background knowledge on the study of oncology or hematology may make better use of the developed summarizer as it is effective in terms of its context. Nonetheless, future researches may further improve on the linguistic aspect of the software. Additionally, the automated text summarizer is currently unable to properly process PDF files for accurate sentence scoring and clean summary output. This resulted in the PDF files having to be manually imported into text files.

To determine whether the summarizer is able to effectively include pertinent data in the treatment of leukemia, experts in the area of hematology or oncology will be gathered and asked to assess the performance of the software. Prospect evaluators include hematologists, who are specialists in the study of blood and blood-related diseases, such as leukemia; oncologists, who are doctors that treat cancer; or hematologist-oncologists, who are physicians that specialize in cancer-related blood diseases. At the present time, the researchers are still on the process of contacting biomedical experts for the further evaluation of the automated text summarizer.

The summarizer should also be evaluated by quantitative means, such as the classical Information Retrieval (IR) measures. The IR measures calculate for the following: precision, the ratio of the number of extracted or abstracted sentences that are actually correct over the total number of extracted or abstracted sentences; recall, the ratio of the number of extracted or abstracted sentences that are actually correct over the total number of correct sentences; F-score, a composite measure that combines precision and recall into one value. These evaluation measures can be implemented by having experts in hematology or oncology create gold standard summaries that are to be compared with the software-generated summaries.

6. REFERENCES

[1] Abstracts. nd. The Writing Center. Retrieved from https://writing.wisc.edu/Handbook/presentations_abstracts.html

[2] Afantenos, S., Karkaletsis, V., and Stamatopoulos, P. 2005. Summarization from medical documents: A survey. Journal of Artificial Intelligence in Medicine, 33, 2 (Feb. 2005), 157-177. DOI: https://doi.org/10.1016/j.artmed.2004.07.017

[3] Aramaki, E., Miura, Y., Tonoike, M., Tomoko, O., Mashuichi, H., and Ohe, K. 2009. Text2Table: Medical text summarization system based on named entity recognition and modality identification. In Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing. ACL, Stroudsburg, PA, 185-192.

[4] Basagic, R., Krupic, D., and Suzic, B. 2009. Automatic text summarization. Information Search and Retrieval, WS, (2009).

[5] Bird, S. and Loper, E. 2002. NLTK: the Natural Language Toolkit. In ETMTNLP '02 Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics - Volume 1. Stroudsburg, Pennsylvania, USA, 63-70. DOI: 10.3115/1118108.1118117

[6] Cancer Research UK. 2015. What is cancer? (December 2015). Retrieved August 24, 2017 from http://www.cancerresearchuk.org/about-cancer/cancer-symptoms/why-is-early-diagnosis-important

[7] Cancer Research UK. 2015. Why is early diagnosis important? (April 2015). Retrieved August 24, 2017 from http://www.cancerresearchuk.org/about-cancer/cancer-symptoms/why-is-early-diagnosis-important

[8] Cancers that develop in children. 2016, August 22. American Cancer Society. Retrieved August 24, 2017, from https://www.cancer.org/cancer/cancer-in-children/types-of-childhood-cancers.html

[9] Cao, Y. G., Liu, F., Simpson, P., Antieau, L., Bennett, A., Cimino, J. J., Ely, J., and Yu, H. 2011. AskHERMES: An online question answering system for complex clinical questions. Journal of biomedical informatics, 44, 2 (Apr. 2011), 277-288. DOI: https://doi.org/10.1016/j.jbi.2011.01.004

[10] Chowdhury, G. 2005. Natural language processing. Annual Review of Information and Technology. 31, 1 (31, Jan. 2005), 51-89. DOI: 10.1002/aris.1440370103

[11] Chuang, W. T. and Yang, J. 2000. Extracting sentence segments for text summarization: a machine learning approach. In Proceedings of the 23rd Annual international ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, New York, NY, 152-159. DOI: 10.1145/345508.345566

[12] Clough, P. and Sanderson, M. 2013. Evaluating the performance of information retrieval systems using test collections. Information Research, 18, 2 (Jun. 2013), 1-10.

[13] Coulter, C. and Ellins, J. Effectiveness of strategies for informing, educating, and involving patients. BMJ. 335, 7609 (May 2007), 24–27.

[14] Crosta, P. 2015. Cancer: Facts, causes, symptoms and research. Retrieved August 24, 2017, from http://www.medicalnewstoday.com/info/cancer-oncology

[15] Driscoll, D. and Kasztalska, A. 2013. Writing the experimental report: Methods, results, and discussion. Retrieved from https://owl.english.purdue.edu/owl/resource/670/04/

[16] Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D., and Bray, F. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012 International Journal of Cancer 136, 5 (March 2015), E359–E386.

[17] Ferreira, R., de Souza Cabral, L., Lins, R. D., Silva, G. P., Freitas, F., Cavalcanti, G., Lima, R., Simske, S. J., and Favaro, L. Assessing sentence scoring techniques for extractive text summarization Expert Systems with Applications 40, 14 (Oct. 2013), 5755-5764.

[18] Gupta, V., and Lehal, G. S. 2010. A survey of text summarization extractive techniques Journal of Emerging Technologies in Web Intelligence 2, 3 (Aug. 2010), 258-268.

[19] Hanahan, D. and Weinberg, R.A. 2000. The hallmarks of cancer. Cell. 100, 1 (2000), 57-70. DOI: https://doi.org/10.1016/S0092-8674(00)81683-9

[20] Hassel, M. 2004. Evaluation of automatic text summarization. Licentiate Thesis (2004), 1–75.

[21] Hovy, E. and Yew, C. 1998, October. Automated text summarization and the SUMMARIST system. Proceedings of a workshop on held at Baltimore, Maryland. (13-15, Oct. 1998), 197-214. DOI: 10.3115/1119089.1119121

[22] ILO. International Labour Organization: Summaries and executive summaries. (2005). Retrieved February 15, 2018 from

http://www.colelearning.net/ilo/English/Module_2b/038_key_elements.htm
[23] Jemal, A., Bray, F., Center, M., Ferlay, Ward, E., and Forman, D. Global cancer statistics CA: A Cancer Journal for Clinicians 61. 2 (Apr. 2011), 69-90. DOI: 10.3322/caac.20107
[24] Josef Steinberger and Karel Je˘zek. Evaluation measures for text summarization. Computing and Informatics 28, 2 (Jan. 2012), 1001-1026.
[25] Key elements of the research proposal. 2010. Retrieved from http://www.bcps.org/offices/lis/researchcourse/key_elements.html
[26] Kolata, G. 2009, April 23. Advances elusive in the drive to cure cancer. Retrieved from http://www.nytimes.com/2009/04/24/health/policy/24cancer.html
[27] Leukemia - treatment overview. n.d. Retrieved August 24, 2017 from http://www.webmd.com/cancer/tc/leukemia-treatment-overview#1
[28] Lloret, E. and Palomar, M. Challenging issues of automatic summarization: relevance detection and quality-based evaluation. Informatica. 34, 1 (2010).
[29] Managing cancer as a chronic illness. 2016, February 12. Retrieved August 24, 2017 from https://wuww.cancer.org/treatment/survivorship-during-and-after-treatment/when-cancer-doesnt-go-away.html#written_by
[30] Mani, I. and Maybury, M. Advances in Automated Text Summarization, Massachusetts Institute of Technology, London.
[31] McCowan, I. A., Moore, D., Dines, J., Gatica-Perez, D., Flynn, M., Wellner, P., and Bourlard, H. 2004. On the use of information retrieval measures for speech recognition evaluation. IDIAP. Martigny, Switzerland.
[32] Meena, W. K. and Gopalani, D. 2014. Analysis of Sentence Scoring Methods for Extractive Automatic Text Summarization. In Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies. Udaipur, Rajasthan, India, 1-6.
[33] Menon, D. 2016. 7 trusted medical journal search engines. (June 2016). Retrieved from https://www.healthwriterhub.com/medical-journal-search-engines/
[34] Mishra, R., Kumar, P., and Bhasker. B. A web recommendation system considering sequential information. Decision Support Systems 75 (2015), 1-10.
[35] Moradi, M. and Ghadiri, N. Quantifying the informativeness for biomedical literature summarization: An itemset mining method. Computer Methods and Programs in Biomedicine 146 (2017), 77-89.
[36] Neto, J., Freitas, A., and Kaestner, C. 2002. Automatic text summarization using a machine learning approach. In Brazilian Symposium on Artificial Intelligence. Springer, Berlin, Heidelberg, 205-215.
[37] Pustejovsky, P. and Stubbs, A. 2012. Natural Language Annotation for Machine Learning. O'Reilly, Sebastopol, CA.
[38] Reeves, L., Han, H., and Brooks, A. The use of domain-specific concepts in biomedical text summarization. Information Processing & Management 43, 6 (2007), 1765-1776.
[39] Sarkar, K. Using domain knowledge for text summarization in medical domain. International Journal of Recent Trends in Engineering 1, 1 (2009), 200-205.
[40] Sarkar, K., Nasipuri, M., and Ghose, S. Using machine learning for medical document summarization. International Journal of Database Theory and Application 4, 1 (Mar. 2011), 31-48.
[41] Saziyabegum, S. and Sajja, P. S. Review on text summarization evaluation methods. Indian Journal of Computer Science and Engineering. 8, 4 (2017), 497-500.
[42] Solis, C. J., Siy, J. T., Tabirao, E., and Ong, E. 2009. Planning author and character goals for story generation. In Proceedings of the Workshop on Computational Approaches to Linguistic Creativity. Association for Computational Linguistics, 63-70.
[43] WHO. 2017. The top 10 causes of death. (January 2017). Retrieved August 24, 2017 from http://www.who.int/mediacentre/factsheets/fs310/en/
[44] Winzelberg, A., Classen, C., Alpers, G., Roberts, H., Koopman, C., Adams, R., Ernst, H., Dev, P., and Taylor, B. 2003. Evaluation of an internet support group for women with primary breast cancer. Cancer 97, 5 (2003), 1164 - 1173.