Advances in Computer, Signals and Systems (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2023.070903
ISSN 2371-8838 Vol. 7 Num. 9

Research on NLP Based Automatic Summarization Generation Method for Medical Texts

Yuhang Tang

Xi'an Jiaotong-Liverpool University Affiliated School, Beijing, 100086, China

Keywords: Natural Language Processing; Medical texts; Automatic abstract generation method

Abstract: The fundamental concept underpinning text summarization technology revolves around the capacity to encapsulate the original information into a succinct form, thus equipping individuals to promptly extract essential content from vast data repositories and liberating users from the cumbersome task of processing extensive textual material. In recent years, the exponential proliferation of data in biomedical literature, patient case records, and healthcare documentation, has presented a pressing challenge. This research undertakes the integration of Natural Language Processing (NLP)-related technologies into the domain of medical text summarization. It puts forth a novel solution for generative automatic summarization, with a specific focus on enhancing the model's proficiency in assimilating the semantic nuances inherent in biomedical texts. The methodology incorporates within existing text summarization frameworks to optimize the model's efficacy in handling biomedical data. The empirical findings presented in this study attest to the remarkable precision of the sentence similarity calculation method introduced herein. In a comparative analysis against four alternative methodologies, this approach achieves a high accuracy rate of 90.6%. This outcome highlights the superior predictive performance of the sentence integration similarity calculation method proposed in this research.

1. Introduction

With the advancement in electronic technology, the quantity of textual information has surpassed the threshold of manual handling. Consequently, the pursuit of natural language processing (NLP) and automatic summarization technology have gained paramount importance and urgency. This technology plays a pivotal role in alleviating the burdensome effects of information overload, effectively hindering the utility of data. The fundamental premise of text summarization technology involves the condensation of extensive original textual content into succinct language, thereby enabling the application of models to expedite the extraction of pivotal content from copious information, thus liberating users from the onerous task of processing voluminous text [1]. Currently, automatic summarization technology has manifested its utility across a multitude of domains, notably in the realms of public opinion analysis and information retrieval. Its potential applications span a wide spectrum, including sentiment analysis, recommendation systems and content production. Particularly within the purview of information security, the direct processing of user generated data on social media platform by public opinion monitoring systems exerts a

substantial computational burden. By preserving the core information in its original form and compressing it before submission to the monitoring system, computational demands can be judiciously mitigated [2-3].

In the sphere of biomedicine, various electronic resources have emerged, such as online literature databases and electronic health record systems, to support clinical practitioners and researchers in managing information [4]. In recent years, data volumes have experienced exponential growth in domains such as biomedical literature, patient cases, health records, and similar contexts. This article endeavors to amalgamate NLP-related technologies into the challenge of automatic medical text summarization, proposing an innovative solution for generative automatic summarization to enhance its quality and applicability within the domain of core information retrieval [5]. Given that a substantial proportion of biological information is documented in textual form, text mining technology is frequently employed in biomedical research. Text mining requires techniques in several domains such as data structuring, information retrieval, and NLP [6]. Leveraging NLP technology, valuable information can be effectively unearthed. Numerous researchers have harnessed text mining tools to extract valuable knowledge for human benefit. This article integrates NLP technology with existing text summarization methods to enhance the model's performance when handling biomedical texts [7]. Furthermore, to mitigate the adverse effects of redundant background information on text summary models, this article advocates the use of NLP to enhance the comprehension of generative text summaries. This enhancement equips the model to comprehensively capture the information contained in the entire text and produce higher-quality abstracts.

2. Research on Automatic Abstract Generation Method for Medical Text

2.1. Model Introduction

Abstract generation models can be broadly categorized into two types: extraction-based models and abstract-based models. Extraction models identify important phrases or sentences within the input document based on predetermined parameters and reassemble them to create a new summary. In extraction-based text summarization, the primary challenge lies in selecting sentences that best represent the entire document. Common techniques for extraction-based text summarization include topic-based sentence extraction, which assesses sentence relevance based on specific topics of interest, and centrality-based sentence extraction, which chooses sentences closely related to others, assuming they provide the most comprehensive coverage of the document [8].

On the other hand, abstract-based models require the use of natural language generation techniques to rephrase sentences. However, these models are not as frequently employed as extraction-based techniques. Common abstract-based techniques include object structuring and sentence semantic understanding. Most object structuring techniques rely on tree structures, entities, and rules. Although they can yield favorable results in certain instances, they may face challenges related to structural scalability and large-scale generation tasks [9]. The process of abstract generation typically involves segmenting the original text information into words and conducting word vectorization processing. This process encompasses calculating part-of-speech, word frequency, and inverse text frequency, ultimately resulting in a word vector sequence that serves as input for the subsequent stage [10]. Subsequently, the surrounding vocabulary of high-frequency terms in the corpus is tallied to create a neighboring vocabulary table, which assists in forming the decoder's vocabulary table. Extensive training on substantial medical text data allows the model to acquire contextual semantic representations. Often, this model is fine-tuned as a pre-training model and applied to various natural language processing tasks. In text summarization tasks, the pre-trained model can function as an effective feature extractor, enabling the model to comprehensively

comprehend the entire text information. The word vector sequence from the previous stage is sequentially input into the encoder. At each step, the encoder generates a semantic vector for the current time step. These semantic vectors are then amalgamated to produce the semantic vector for the entire text, which is then passed on to the subsequent stage. Subsequently, the NLP output is channeled into an LSTM, a commonly used component in sequence-to-sequence generation tasks. It serves the purpose of learning the features of medical abstracts and the characteristics of sequences, determining whether sentences meet the criteria for being considered abstracts. It has been determined that employing NLP as the abstract extractor produced the most favorable results.

2.2. Summary generation

An influential determinant in the realm of text summarization pertains to the method employed for text representation. A proficient text representation methodology serves to enhance the model's capacity for comprehensive comprehension of the complete textual content, thereby ameliorating the overall model performance. The NLP abstract generation model is meticulously trained through an extensive corpus dataset, ultimately affording it a context-dependent textual representation. This specialized representation technique has demonstrated remarkable efficacy in numerous natural language processing tasks. The integral components of the full-text feature primarily encompass alias features, among other constituents.

The alias features function as repositories for designated entities identified in the preceding sentence of the text, maintaining them in a structured list. Whenever the system encounters a candidate word potentially denoting a named entity, the named entity alias recognition algorithm is activated, dynamically ascertaining whether the candidate word indeed serves as an alias for any entity previously listed. This procedure is executed through the integration of NLP technology, which interprets the resemblance between words or sentences as a form of recommendation relationship. Additionally, the significance of nodes is not solely contingent on the count of votes and edges but extends to encompass node importance. This holistic approach contributes to the determination of the score or rank assigned to each vertex in the network. A graphical depiction of this comprehensive method is elucidated in Figure 1.

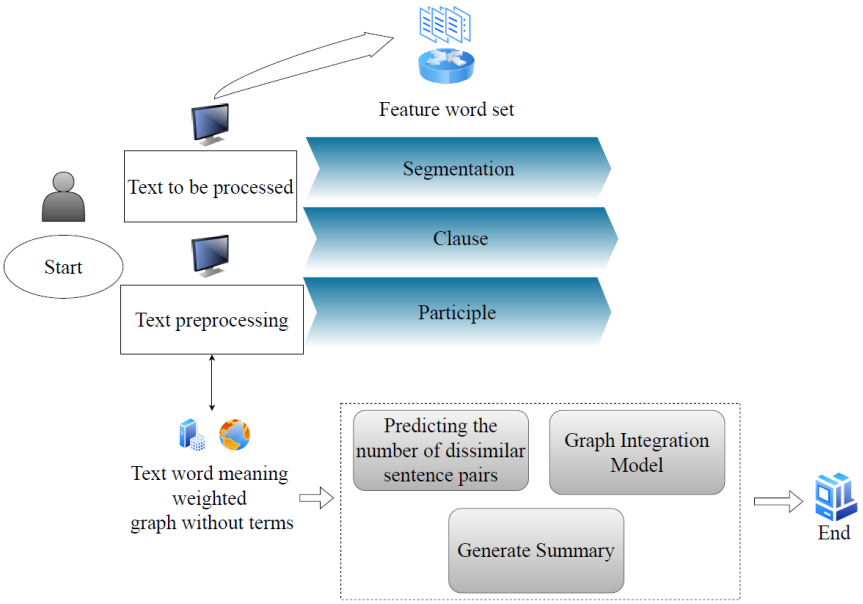

Figure 1: Process of Medical Text Automatic Summary Generation Method Based on NLP

After obtaining the text integration similarity graph through the graph integration model, the

NLP iterative convergence method is used to sort the graph nodes. The iterative calculation formula for sentence node weight is:

$$WS(V_t)=1-d+d\times\sum_{V_j\in in(V_i)}\frac{w_{ij}}{\sum_{V_k\in out(V_j)}w_{jk}} \tag{1}$$

Where $WS(V_t)$ is the weight value of node V_i, d is the damping coefficient, which is generally 0.85. In the figure, the probability of a node jumping to another node is $(1-d)$, $in(V_i)$ is the set of all nodes that $out(V_j)$ points to node V_i, indicating the set of all nodes that node V_j points to, and W_{ij} is the weight of the edge between node V_i and node V_j.

Errors in the automatic text summarization process are occasionally attributed to boundary recognition errors, which tend to emerge primarily due to the prevalence of multi-word phrases within named entities. Boundary recognition errors typically occur due to inadequate recognition of both the left and right boundaries, frequently leading to the inadvertent omission of nouns, adjectives, and adverbs with less conspicuous attributes. To address this challenge, advanced NLP techniques facilitate an intricate examination of the identified text during the automatic summarization process. This scrutiny is complemented by the implementation of heuristic rules and probability statistics, which serve the purpose of expanding and rectifying the boundaries pertaining to the identified named entities.

3. Result analysis

3.1. Sentence integration similarity evaluation

In the context of biomedical texts, the evaluation of similarity operates at a higher conceptual level, transcending mere word and sentence-level comparisons. To illustrate this concept, consider the terms "fit" and "seizure," which may appear dissimilar from a non-medical context, but in fact share a closely related underlying medical pathology. Such subtle nuances are prevalent in clinical records, underscoring the significance of equipping the model with a nuanced understanding of background concepts inherent to terms. This comprehension is pivotal for the text abstraction model to effectively decipher the intricate semantics embedded within biomedical texts.

The dataset utilized for this research encompasses an English corpus specifically designed for manual sentence similarity assessment. It comprises a substantial total of 580,000 sentence pairs, which are further subdivided into 650,857 pairs allocated for training and an additional 10,000 pairs for testing. To facilitate the assessment of sentence similarity, three distinct labels—namely "entailment," "contradiction," and "neutral"—are assigned upon the logical relationship established between reasoning premises and assumptions. In the context of this experiment, sentence pairs designated as "entailment" are selected to denote semantically similar samples, while those categorized as "contradiction" are indicative of semantically dissimilar samples. Consequently, this framework yields a comprehensive training set encompassing 486,648 sentence pairs and an intricate test set containing 7,504 sentence pairs. The results of this analysis, conducted with a specific focus on the comparison with other methodologies such as the TextRank method, TF-IDF, and K-Means, employing a predetermined similarity threshold of 0.6, are thoughtfully presented in Figure 2.

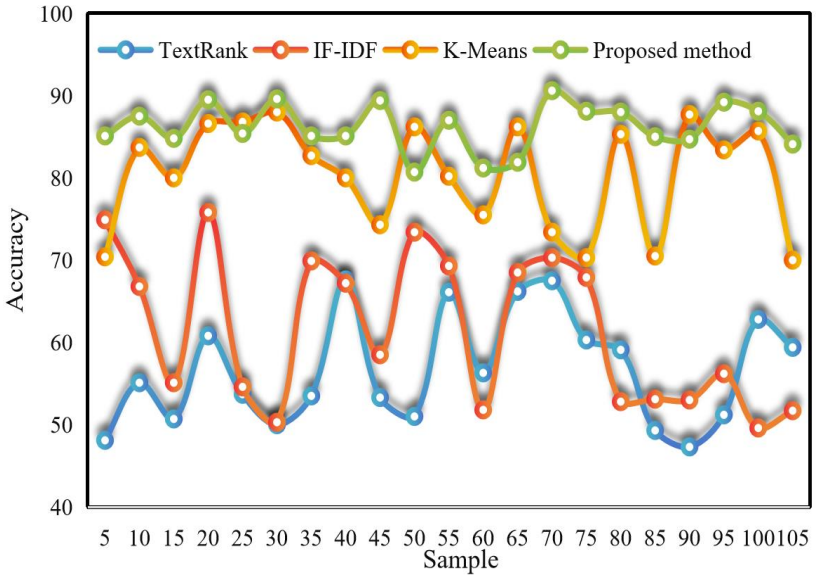

Figure 2: Accuracy

From the comparison of experimental results in Figure 2, it can be observed that the NLP algorithm used in this paper to calculate sentence similarity exhibits a high accuracy in comparison with the four alternative methods, with an accuracy rate of up to 90.6%. These results demonstrate the superior predictive performance of the method, affirming its capabilities in the realm of forecasting sentence similarity.

3.2. Summary evaluation

This section outlines a method for generating single-document abstracts of biomedical articles. The approach entails the extraction of semantic maps denoting concepts and relationships present in the document, facilitated by the utilization of a comprehensive super thesaurus and a semantic network. Subsequently, a clustering algorithm grounded in degree centrality is employed to discern salient text topics, and summary sentences are extracted based on the identification of distinct topics within each sentence. However, it is pertinent to acknowledge that this method predominantly relies on word frequency and does not encompass the incorporation of semantic information pertaining to the vocabulary. Consequently, it may consider numerous proprietary nouns as independent entities. For instance, consider deoxyribonucleotides, which are basic constituent units exhibiting a strong interrelation; such associations might be inadvertently overlooked within the framework of this model.

The primary focus of the research presented in this article is the development of a single-document automatic summary generation technology tailored for lengthy texts. The experimentation employs the CNN/DailyMail dataset for single-text automatic text summarization. The dataset comprises a training set consisting of 296,853 pairs, a validation set encompassing 17,346 pairs, and a test set comprising 13,451 pairs. The training set serves as the basis for summary evaluation. On average, the training set encompasses approximately 854 words and about 29 sentences, while the corresponding summaries are characterized by an average of 56 words and approximately 3.86 sentences.

To assess the proposed automatic summarization method, which is rooted in a graph ensemble model, a comparative analysis is conducted against conventional text summarization extraction methods, namely TextRank, TF-IDF, and K-Means. The results of this comparative evaluation are presented in Table 1.

Table 1: Score Results of Evaluation Indicators for Each Method

Method	ROUGE-1	ROUGE-2	ROUGE-3
TextRank	31.2	15.4	30.1
IF-IDF	36.2	18.4	33.5
K-Means	35.7	16.1	32.7
Proposed method	38.9	20.3	36.4

By comparing the experimental results presented in Table 1, it is evident that the proposed method in this study exhibits substantial prowess in managing lengthy textual inputs, with all evaluation metrics returning commendable outcomes. Notably, when contrasted with the TextRank method and the evaluation metrics predicated on TF-IDF and K-Means, there emerges a noteworthy performance enhancement of approximately 10%. It is noteworthy that the TextRank method, as well as the generative summary methods grounded in TF-IDF and K-Means, are better suited for handling concise textual inputs. In the context of protracted textual sequences, the encoder may grapple with the accurate extraction of semantic information due to the inherent challenges associated with long-distance dependencies. Such challenges can impede the effective convergence of the model, consequently compromising the precision of summary generation.

The scores achieved across the spectrum of experimental evaluation metrics markedly surpass those attained by other methodologies, thereby underscoring the preeminence of the abstract generation technology proposed in this paper.

4. Conclusions

Automatic summarization has been a longstanding subject of research within the realm of NLP. In recent times, generative automatic summarization techniques have garnered substantial attention as a focal point of research. In the domain of text summarization technology, the central challenge has perennially revolved around enhancing the model's ability to comprehensively comprehend textual information and subsequently extract the core content for the generation of high-quality summaries. This paper is dedicated to the augmentation of text summarization technology, with a particular emphasis on its application within the biomedical domain.

Within the scope of extractive text summarization technology, the methodology primarily entails the evaluation of the original sentences' importance, subsequently selecting the highest-ranked sentences as the central content of the text. These chosen sentences are then mapped with the super thesaurus specific to the biomedical field to derive globally significant term-concept pairs. These concept pairs are subsequently integrated into the model's attention mechanism, thus infusing it with biomedical background knowledge, thereby guiding the model's focus towards key contextual information within the text. This integration is instrumental in effecting the extraction of semantic maps delineating concepts and relationships from the hyperthesaurus embedded within the documents. To achieve this, the hyperthesaurus and a semantic network are actively employed. Furthermore, the process entails the utilization of a clustering algorithm that hinges on degree centrality to pinpoint salient text topics, culminating in the extraction of abstract sentences predicated upon the presence of various identified topics within each sentence.

In summation, the research findings showcase the remarkable precision of the sentence similarity calculation method introduced in this paper. When contrasted with four alternative methods, the proposed method achieves an exceptional accuracy rate of 90.6%. This serves as an evidence for the efficacy of the sentence integration similarity calculation approach outlined in this paper, manifesting as a superior predictor of sentence similarity.

References

[1] Yamamoto S, Fukuhara Y, Suzuki R, et al. Automatic Paper Summary Generation from Visual and Textual Information [J]. Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology, 2021, 10(7):11-14.
[2] Liu Y, Yang Y, Huang Y. Automatic Generation of Comparative Summary for Scientific Literature[J]. International Journal of Performability Engineering, 2018, 14(7):1570-1579.
[3] Xu F, Yi G, Qi W, et al. Automatic summary of short text based on Seq2Seq and keywords correction[J]. Computer Engineering and Design, 2022, 11(9):8-14.
[4] Menger, Vincent, Scheepers, et al. DEDUCE: A pattern matching method for automatic de-identification of Dutch medical text [J]. Telematics and informatics, 2022, 19(7):35-40.
[5] Ozyegen O, Kabe D, Cevik M. Word-level Text Highlighting of Medical Texts forTelehealth Services[J]. Journal of Medical Imaging and Health Informatics, 2021, 11(4):36-40.
[6] Jiamin Cao, Jiawei Wang, et al. Semi-Automatic Synthetic Computed Tomography Generation for Abdomens Using Transfer Learning and Semi-Supervised Classification [J]. Journal of Medical Imaging and Health Informatics, 2019, 9(9):11-1
14.
[7] Xingqiang W, Na M. The Automatic Generation Method of Shared Document of Electronic Medical Records [J]. China Digital Medicine, 2019, 13(4):21-26.
[8] Matentzoglu, N. Ontology-Based Generation of Medical, Multi-Term MCQs. [J]. International Journal of Artificial Intelligence in Education, 2019, 29(5):5-9.
[9] Zhou Q, Peng W, Tang D. Automatic recommendation of medical departments to outpatients based on text analyses and medical knowledge graph[J]. Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology, 2021, 13(2):41-45.
[10] Yashaswini S, Shylaja S S. Metrics for Automatic Evaluation of Text from NLP Models for Text to Scene Generation [J]. European Open Science Publishing, 2021, 12(4):9-12.