Manuscript Submitted	15..6.2023
Accepted	23.12.2023
Published	31.12.2023

Precision in Communication: A Natural Language Processing Approach to Text Summarization

Md Serajun Nabi, Zaidan Mufaddhal, Khai Lin Khant & Ahmed Mansour Raufi
School of Computing and Informatics
Albukhary International University
serajun.nabi@student.aiu.edu.my

Akibu Mahmoud Abdullahi
School of Computing and Informatics
Albukhary International University
akibu.abdullahi@aiu.edu.my

Abstract

This study presents an experimental method to understanding the fundamental ideas of natural language processing and developing a system for automatically summarizing material in a form that users can grasp in a short amount of time. Tokenization, word embedding's, cleaning sentences, deleting stop words, and utilizing Cosine Similarity and Networks to determine similarity between phrases and pick the most essential ones to create the summary are all employed in the experiment. We discovered that advanced techniques, such as machine learning algorithms, may successfully extract essential concepts and create meaningful summaries from a given text. We also discovered that by considering the general subject and arrangement of the text, as well as applying Cosine Similarity and Networks, we could create more accurate and thorough summaries.

Keywords: *Natural Language Processing, Text Summarization, Machine Learning.*

1. Introduction

1.2 Background of Study

Automatic Text Summarization is a method that reduces lengthy texts and creates summaries to convey the intended content (Kumar et al., 2018). It is critical to develop a punctuation process to quickly condense lengthy texts while maintaining their primary points because data is being encoded at an unprecedented rate. Summarizing also expedites information searches, reduces reading time, and provides the most information on a single subject. The basic goal of an electronic text summary is to shorten the reference text into a more manageable size while still maintaining its understanding and value. And the report is described as text that is derived from one or more papers that communicate relevant knowledge in the initial text, which does not exceed half of the primary texts and is typically somewhat more constrained than that. There are two types of text summarization systems: extractive summarization systems and abstractive summarization systems. In order to create summaries, extractive summarization systems copy significant portions of the original text and then mix those portions and sentences. Sentence importance is determined by linguistic and statistical characteristics. On the other hand, the phrases produced by abstractive summarization methods may be rephrased or may contain terms that weren't present in the source text. Naturally abstractive methods are more difficult. To create a great abstractive summary, the model must first fully comprehend the content before attempting to condense that understanding into a few words, maybe utilizing novel language. much more difficult than extractive. possesses sophisticated talents like the ability to generalize, paraphrase, and incorporate

real-world knowledge. Due to the simplicity of developing hard-coded algorithms to select essential phrases rather than generating new ones, the majority of the effort has typically been on extractive techniques. Additionally, it guarantees a logical, linguistically sound summary. However, because they are so constrained, they frequently don't summarize lengthy and complex texts well (Muqtadir et al., 2020).

1.2 Problem Statement

In our daily lives, we are quite busy with many other significant things. Sometimes, we do not have enough time to read a long text, even though it is important. In addition, after reading a lengthy text, we were perplexed as to what the text's main point was. To solve this problem, we are going to summarize the large text into a short summary, which will reduce the user's reading time.

1.3 Objectives

- To understand the main concepts of natural language processing.
- To summarize the text automatically.
- To make it easier for users to understand the summary of a text in a short period of time.

2. Literature Review

2.1 Related Work

According to Islam Talukder et al., (2020), entitled "Abstractive Text Summarization. The two types of text summarization techniques are called abstractive and extractive. It generates a summary that resembles a human's, abstractive is superior to extractive. As different abstractive approaches, the Word Graph Methodology, Semantic Graph Reduction Algorithm, and Markov Clustering Principle were contrasted in this work. They contrasted them based on the methods used to remove redundancy, the degree to which the text was reduced, and the degree to which it was semantically and syntactically correct.

According to Pratibha Devi Hosur et al. (2017), the system suggests using unsupervised learning while doing automatic text summarization. The input text document, pre-processing, the Lesk approach, and eventually the creation of the summaries are all covered in-depth in this paper's review of text summarization using NLP. the Lesk algorithm's findings, calculations, analysis, and recommended system.

Hovy and Lin (1998) presented a typology of summaries regarding the traits of the source text(s), the traits of generated summaries, and the goal of summarization. The authors suggested two different forms of summaries, indicative and informative, under the "use" category of the typology. Without revealing its contents, a suggestive summary seeks to convey the main ideas of the supplied text(s). The goal of an informative summary, on the other hand, is to accurately reflect (some of) the content of a text document and to compel the reader to explain (some of) its contents. Referring back to our issue, inquiries might be thought of as "indicative" summaries. After reading the questions, the reader is required to explain the topic of the input text, not necessarily what was written in it.

According to Ravali Boorugu and Kumar et al., (2018) in their review, some of the outstanding achievements in the field of text summarization were thoroughly detailed. They also covered several methods of text summarization. The three primary types of text summaries are listed below. And they are, depending on the output type, the purpose, and the type of input. Product reviews in an online market can be summarized using text summarizing algorithms. This can make it easier for buyers to read lengthy evaluations. The survey in this research comes to the conclusion that the Seq2Seq model, combined with the LSTM and attention mechanism, can be used to summarize online product reviews more accurately.

According to Y. Du, and Hua (2020), they proposed a text summarization method that can be used for summarizing news based on Multi-Feature and Fuzzy Logic. This work primarily focuses on four areas, including using NLTK to pre-process the news and extracting textual features like word features, word frequencies, word properties, and so on. The next step is to use the heuristic search method genetic algorithm to apply weights to the extracted news feature, and then use fuzzy logic to award scores to the phrases. Fuzzy logic mimics the concept assessment and reasoning processes of the human brain. Additionally, the proposed method is contrasted in this work with others, including MS Word, GCD, SOM, System 19, System 311, SDS-NNGA System 2121, and Ranking SVM. And the suggested approach outperforms alternative approaches.

According to G. Vijay Kumar and V. Valli Kumari (2012), regular patterns can help text summarizing algorithms extract relevant keywords. They mostly talked about the approaches of the abstractive and extractive methods. Neelima Bhatia et al., (2017) reviewed the well-known and significant effort made in the field of units and numerous archive outlines. The authors looked at method-based approaches to text summarization. Theme-based techniques, talk-based methodology, term-based recurrence strategy, diagram-based technique, time-based strategy, division and combining strategy, semantic dependence strategy, and latent semantic analysis These technique-based strategies include approaches that are semantically based, lexically chained, and reliant on flimsy justification.

According to Shohreh Rad Rahimi et al. (2017), NLP investigations are more interested in summarizing literary data. A text summary is described by the authors of this study as an interplay between naturally condensing the kind of a given report and holding its data content source into a more condensed variant with the proper importance. The authors of this research also outlined the relationship between text mining and text outlines. Finally, this study looks at various approaches to dealing with text outlines, including statistical methodologies, lexical chain-based methodologies, cluster-based methodologies, and fuzzy logic-based methodologies.

Luhn et al. (1958) developed a technique to extract important sentences from the text using features like word and phrase frequency . They suggested discarding very high- frequency common terms and weighing sentences in a document according to high-frequency words.
Edmundson et al. (1969) proposed a paradigm based on key phrases that used the following three techniques to calculate sentence weight in addition to the usual frequency-based weights:

- Cue Method: Using the presence or absence of specific cue words from the cue dictionary, the relevance of a sentence is determined.
- Title Method: The weight of a sentence is determined by adding together all the content words found in the text's title and headers.
- Location Method: This approach makes the assumption that sentences that appear at the start of a document as well as the start of each paragraph are more likely to be meaningful.

3. Methodology

3.1 Method

In order to summarize our material, an extraction approach based on conventional and straightforward algorithms are selected. For instance, we keep track of all the key words and their frequencies in the dictionary when we want to summarize our text using the frequency technique. We store the sentences that contain a high frequency word in our final summary based on their usage. This indicates that the words in our summary attest to their inclusion in the original text.

3.2 Techniques

The experiments' technique includes five major steps:

3.2.1 Tokenization:
Tokenization technique was used in this research project, the process tokenization is to breaking down input material into individual words or phrases known as tokens. Tokenization is an important step in natural language processing since it allows the system to recognize and deal with the text's various components.

3.2.2 Word Embedding:
After tokenizing the text, word embeddings were employed to represent the words numerically. Word embeddings are a method of encoding words in a high-dimensional space, with semantically comparable words clustered together. To increase the performance of our models, we employed pre-trained word embedding's.

3.2.3 Deleting Superfluous:
We next preprocessed the text by deleting any superfluous or redundant information, such as special characters, numerals, and punctuation marks. We also deleted stop words, which are frequent terms that have no significance, such as "the," "is," "and," and so on.

3.2.4 Stop Words:
Stop words were eliminated from the text since they add little sense and make the text more difficult to interpret.

3.2.5 Cosine Similarity:
Cosine similarity was utilized in the research to compare the similarity of all sentences in a given text. Then, based on the highest similarity scores, we chose the most informative and significant lines and included them in the summary.

3.2.6 Results and comparison:
The performance was compared of the models using numerous assessment measures, including the Rouge score and the F1-score. We also compared the findings of our models to those of other cutting-edge models to evaluate how our method fared.

3.3 Dataset
For the dataset, there are 3 columns which are article-id, article, and source. For this row, there are 7 rows, which means there are 7 articles crossing the internet.

4. Result
We described in the methodology a thorough strategy for text summarization that makes use of natural language processing methods. Tokenization, word embedding, stop and unnecessary word removal, and cosine similarity computation were the processes involved. Notably, in order to create a coherent summary, we used the extractive approach, which chooses sentences straight out of the document based on a scoring system.

We closely adhered to the specified methodology in order to get into the details of result generation. We employed a methodical procedure to find and eliminate unnecessary and stop words after tokenizing the text and embedding words in order to extract the main ideas. We were then able to assess each sentence's relevance and significance by using cosine similarity.

```
ARTICLE:
Maria Sharapova has basically no friends as tennis players on the WTA Tour. The Russian player has no problems in openly spe
aking about it and in a recent interview she said: 'I don't really hide any feelings too much. I think everyone knows this i
s my job here. When I'm on the courts or when I'm on the court playing, I'm a competitor and I want to beat every single per
son whether they're in the locker room or across the net.So I'm not the one to strike up a conversation about the weather an
d know that in the next few minutes I have to go and try to win a tennis match. I'm a pretty competitive girl. I say my hell
os, but I'm not sending any players flowers as well. Uhm, I'm not really friendly or close to many players. I have not a lot
of friends away from the courts.' When she said she is not really close to a lot of players, is that something strategic tha
t she is doing? Is it different on the men's tour than the women's tour? 'No, not at all. I think just because you're in the
same sport doesn't mean that you have to be friends with everyone just because you're categorized, you're a tennis player, s
o you're going to get along with tennis players. I think every person has different interests. I have friends that have comp
letely different jobs and interests, and I've met them in very different parts of my life. I think everyone just thinks beca
use we're tennis players we should be the greatest of friends. But ultimately tennis is just a very small part of what we d
o. There are so many other things that we're interested in, that we do


SUMMARY:
When I'm on the courts or when I'm on the court playing, I'm a competitor and I want to beat every single person whether the
y're in the locker room or across the net.So I'm not the one to strike up a conversation about the weather and know that in
the next few minutes I have to go and try to win a tennis match.
```

Figure 1: Output of the article summarization_1

```
ARTICLE:
BASEL, Switzerland (AP), Roger Federer advanced to the 14th Swiss Indoors final of his career by beating seventh-seeded Dani
il Medvedev 6-1, 6-4 on Saturday. Seeking a ninth title at his hometown event, and a 99th overall, Federer will play 93th-ra
nked Marius Copil on Sunday. Federer dominated the 20th-ranked Medvedev and had his first match-point chance to break serve
again at 5-1. He then dropped his serve to love, and let another match point slip in Medvedev's next service game by netting
a backhand. He clinched on his fourth chance when Medvedev netted from the baseline. Copil upset expectations of a Federer f
inal against Alexander Zverev in a 6-3, 6-7 (6), 6-4 win over the fifth-ranked German in the earlier semifinal. The Romanian
aims for a first title after arriving at Basel without a career win over a top-10 opponent. Copil has two after also beating
No. 6 Marin Cilic in the second round. Copil fired 26 aces past Zverev and never dropped serve, clinching after 2 1/2 hours
with a forehand volley winner to break Zverev for the second time in the semifinal. He came through two rounds of qualifying
last weekend to reach the Basel main draw, including beating Zverev's older brother, Mischa. Federer had an easier time than
in his only previous match against Medvedev, a three-setter at Shanghai two weeks ago.


SUMMARY:
Major players feel that a big event in late November combined with one in January before the Australian Open will mean too m
uch tennis and too little rest.
```

Figure 2: Output of the article summarization_2

The methodology's discussion of the extractive approach was crucial to the outcomes attained. Sentences were ranked in order of significance by the scoring function used during extraction, which helped to create a concise and logical summary. It is imperative to observe that the steps described in the methodology are in perfect harmony with the actions carried out in the results section, guaranteeing the dependability and coherence of our methodology.

5. Discussion

The experiments we were designed to help us grasp the fundamental ideas of natural language processing and to create a system for automatically summarizing information in a manner that people can understand in a short amount of time.

In this research project, the sophisticated approaches have been discovered such as machine learning algorithms may efficiently extract essential concepts and provide useful summaries of a given text. We discovered that integrating techniques like text compression and sentence extraction can increase the efficiency and efficacy of the summarizing process even more.

To comprehend and summarize text in our tests, we employed a variety of natural language processing approaches. Tokenization, word embedding’s, cleaning sentences, and removing stop words were crucial elements in our process. Tokenization is a crucial stage in natural language processing in which the input text is broken down into individual words or phrases, also known as tokens. This enables the system to comprehend and interact with the text's constituent components.

After tokenizing the text, we employed word embeddings to represent the words numerically. Word embeddings are a method of encoding words in a high-dimensional space, with semantically comparable words clustered together. To increase the performance of our models, we employed pre-trained word embedding's. In addition, we preprocessed the text by deleting any unnecessary or superfluous information such as special characters, digits, and punctuation marks. We also deleted stop words, which are frequent terms that have no significance, such as "the," "is," "and," and so on.

In the experiments, the Cosine Similarity was utilized to discover similarities between sentences in the text in order to construct summaries. The technique measures the similarity between multiple sentences in the text using Cosine Similarity. The sentences with the greatest similarity are then chosen as the most relevant and significant sentence, and they are included in the summary. This method is based on the assumption that semantically related words would have a high degree of similarity when represented as vectors in a high-dimensional space.

Furthermore, Network has been utilized to create a graph of the sentences in a given text, with the edges representing the similarity between sentences as estimated using Cosine Similarity. The most essential sentences in the graph were then identified using centrality metrics such as PageRank and included in the summary. We were able to develop summaries that effectively represented the primary thoughts and ideas offered in the original text by using Network to form a graph of the sentences in the text and select the most relevant sentences.

One of the most important lessons that found from the experiments was the importance of considering the context and structure of the text when creating a summary. We discovered that by considering the general subject and arrangement of the text, we could create more accurate and thorough summaries. Overall, our experiments shown that natural language processing may be utilized to create accurate and useful text summaries, allowing users to grasp the essential ideas in a short period of time.

6. Conclusion

An experiment was carried out in order to grasp the fundamental ideas of natural language processing and to design a system for automatically summarizing information in a manner that users can understand in a short amount of time. Tokenization, word embeddings, cleaning sentences, and removing stop words were all part of the experiment's technique. Additionally, we applied the cosine similarity technique to construct similarities between sentences, then showed them in a graph using the Networkx Python library. The experiment's findings demonstrated that natural language processing may be utilized to create accurate and useful text summaries, allowing users to grasp the important ideas in a short period of time. The investigation also discovered that sophisticated approaches, such as machine learning algorithms, may efficiently extract essential concepts from a given text and provide meaningful summaries.

Acknowledgement

We would like to extend our heartfelt gratitude to Dr. Akibu Mahmoud Abdullahi for his exceptional mentorship and unwavering support as our research supervisor. His invaluable guidance has been instrumental in shaping the outcome of this project. Furthermore, we are immensely thankful to the participants who generously dedicated their time and shared their valuable insights, without whom this study would not have been possible. Lastly, we acknowledge the support provided by our educational institution, which played a vital role in providing the necessary resources and conducive environment for conducting this research.

References

Muqtadir, S., Hussaini, U., Mohd Khan, F., Khan, F., Subhan, A., Scholar, U., & Professor, A. (2020). Text Summarization using Natural Language Processing.

Islam Talukder, Md. A., Abujar, S., Masum, A. K. M., Krishna K P, Y., Dev, S., & V#, K. (2020). Abstractive summarization.

Pratibha Devihosur, Naseer R. (2017) "Automatic Text Summarization Using Natural Language Processing" International Research Journal of Engineering and Technology (IRJET) Volume: 04 Issue: 08.

E. Hovy and C.-Y. Lin, (1998). Automated text summarization and the summarist system. In Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998, pages 197–214. Association for Computational Linguistics.

Kumar, A., Luo, Z., Xu, M., Tufts, S., Sood, A., Aditham, S., Kulathumani, R., & Nguyen, T. (2018). Text Summarization using Natural Language Processing.

G. Vijay Kumar and V. Valli Kumari, (2012), "Sliding Window Technique to Mine Regular Frequent Patterns in Data Streams using Vertical Format", IEEE International Conference on Computational Intelligence and Computing Research.

Potharaju, S. P., & Sreedevi, M. (2017). A Novel Clustering Based Candidate Feature Selection Framework Using Correlation Coefficient for Improving Classification Performance. Journal of Engineering Science & Technology Review, 10(6).

Shohreh Rad Rahimi, Ali Toofan zadeh Mozhdehi and Mohamad Abdolahi, (2017), "An Overview on Extractive Text Summarization"."IEEE 4th International Conference on Knowledge Based Engineering and Innovation" (KBEI), Iran University of Science and Technology – Tehran, Iran.

Hans Peter Luhn. (1958). The automatic creation of literature abstracts. IBM Journal of research and development 2(2), 159–165.

Harold P Edmundson. (1969). New methods in automatic extracting. Journal of the ACM (JACM) 16(2), 264–285.