Summary Generation using Natural Language Processing Techniques and Cosine Similarity

Sayantan Pal[1][0000-0002-5008-0657], Maiga Chang[2][0000-0002-2827-6223], Maria Fernandez Iriarte[2][0000-0002-2597-4288]

[1] Heritage Institute of Technology, Kolkata, WB 700107, India
[2] Athabasca University, Edmonton, AB T5J-3S8, Canada
sayantan.world98@gmail.com, maiga.chang@gmail.com, data@pladio.es

Abstract. The COVID-19 pandemic has led to an unprecedented challenge to public health. It resulted in global efforts to understand, record, and alleviate the disease. This research serves the purpose of generating a relevant summary related to Coronavirus. The research uses the COVID-19 Open Research Dataset (CORD-19) provided by Allen Institute for AI. The dataset contains 236,336 academic full-text articles as of July 19, 2021. This paper introduces a web-based system to handle user questions over the Coronavirus full-text scholarly articles. The system periodically runs backend services to process such large amount article with basic Natural Language Processing (NLP) techniques that include tokenization, N-Grams extraction, and part-of-speech (PoS) tagging. It automatically identifies the keywords from the question and uses cosine similarity to summarize the associated content and present to the user. This research will possibly benefit researchers, health workers as well as other individuals. Moreover, the same service can be used to train with the datasets of different domains (e.g., education) to generate a relevant summary for other user groups (e.g., students).

Keywords: Question & Answering, Information Extraction, Parts of Speech, N-Grams, Coronavirus.

1 Introduction

In the last 18 months, the COVID-19 pandemic has led to a significant change in human lifestyle worldwide. The spread of the disease affects public health, transportation systems, food systems, and everyone's life. Medical institutions and organizations have improved their modes to promote efficient ways to spread awareness among the mass. Spreading awareness and consciousness play a crucial role in decreasing the spread of the outbreak. The research proposes to develop an online summary generation service that would allow users to ask questions related to COVID-19, and in response, they will receive relevant summary to satisfy them.

Researchers publish hundred thousands of articles about Coronavirus include SARS, MERS, SARS-CoV-2 (i.e., COVID-19), and other coronavirus-related topics. We can efficiently utilize the power of computers to process this enormous amount of text data

and extract the essential information from it. The proposed research aims to create a web-based service and system that can identify the keywords from the user's question to generate a relevant summary from those hundred thousand Coronavirus articles. Such system can help medical researchers, health workers, and other individuals to understand the pandemic better.

The proposed system profoundly relies on Natural Processing Techniques to extract the essential information from the pure text articles in the CORD-19 dataset provided by Allen Institute for AI. CORD-19 is a free resource of hundreds of thousands of scholarly articles about Coronaviruses. The system follows a three-step methodology to achieve its goal: (1) extracting and filtering the useful documents from the extensive corpus of compressed scholarly articles; (2) extracting the sentences from the identified documents and Part-of-Speech (POS) tagging the important N-Grams to build the knowledge base; (3) using the processed information stored in the database to generate a relevant summary for the user based on the cosine similarity results of the keyword vectors extracted from both the user's question and the scholarly articles.

This paper is organized as follows: Section 2 starts with the definition of the two involved techniques, Natural Language Processing and Cosine Similarity. Section 3 introduces the way of analyzing and finding a list of Part-of-Speech (PoS) tags for N-Gram, where N is from 1 to 4. Section 3 explains the workflow and methods of the 3-stage summary generation the proposed system uses. Section 5 gives readers an idea of the system as well as the evaluation plan the research team is going to do shortly. At the end Section 6 concludes this research with a summary, the limitations, and the potential future works.

2 Natural Language Processing and Cosine Similarity

2.1 Information Extraction from Text

Information extraction is the task of identifying key phrases and structured information within the text by looking for predefined sequences in the text with pattern matching and natural language processing techniques [1]. It is an essential function in data mining, natural language processing, information retrieval, and Web mining [2]. It has an extensive scope of applications in fields such as Biomedical Literature Mining and Decision support system [3].

One of the primary techniques that information extraction needs is Natural Language Processing (NLP) [4], an area of research that examines the ability of computers to understand a human language. In addition to understanding a language, NLP can be used to extract information from a large corpus of human-readable text. Applications of NLP are extensive [5-7], and it includes several disciplines of study, such as computational linguistics, mathematics, artificial intelligence, machine translation, speech recognition, information retrieval, text processing and, summarization. Various NLP techniques hold enormous potential to convert an unclean corpus to a meaningful text or bag of words [8-10].

A few prominent NLP techniques include Named Entity Recognition (NER), Tokenization, Parts of Speech (PoS) tagging, Stemming and Lemmatization, Natural

Language Generation, and Sentiment Analysis. To process a large corpus of scholarly articles like CORD-19 dataset, Tokenization plays a key role [11]. It is the fundamental step to split comprehensive texts into small tokens that include sentences and words. As tokens are the building blocks of any natural language, the most traditional way of processing the raw text occurs at the token level. Tokens are often grouped and called N-Grams. An N-gram is a contiguous sequence of N tokens from a given sample of text or speech. Tokenization is generally followed by tagging the tokens if necessary [12], and usually PoS tagging is adopted to achieve this. Researchers use them (i.e., Tokenization and PoS tagging) in various tasks such as information retrieval, parsing, Text to Speech (TTS) applications, information extraction, linguistic research for corpora [13-14]. Consider an example sentence "Birds fly high." Table 1 lists the tokenized N-Grams and their PoS for processing the sentence.

Table 1. N-Grams and PoS tags of a given sentence

N value	N-Gram	PoS tag
1	Birds	NNS
1	fly	NN
1	high	JJ
2	Birds fly	NNS-VBP
2	fly high	RB-JJ
3	Birds fly high	NNS-VBP-JJ

2.2 Text Summarization and Cosine Similarity

Text summarization means compressing a considerable section of the source text into a shrunk version that is preserving its knowledge content and overall sense. Text summarization techniques can be extractive and abstractive. An extractive summarization technique requires choosing sentences of a high degree from the document based on word and sentence features and assemble them to generate a summary [15].

Cosine similarity plays a fundamental role in the summarization of text where text-similarity is of primary importance. Cosine similarity measures the similarity between two vectors of an inner product space. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space. The smaller the angle, the higher the cosine similarity. Cosine similarity is often to be used to calculate the similarity among documents in text analysis, irrespective of their size [16]. It is beneficial because even if two similar documents are considerably distant by the Euclidean distance, there is a possibility that they still may be oriented closer collectively.

Summary generation is achieved by representing an attribute of text or the text itself as vectors [17]. To be able to generate these vectors, identification of the keywords is a crucial task. Several methods have been proposed by researchers in the recent past to calculate the similarity between two documents [18-20]. The proposed approach in this paper is different from theirs as we are generating the vectors based on the frequency of the keywords extracted and counted from the scholarly articles in CORD-19 dataset. It ensures a quick generation of the vectors and makes the summarization faster.

3 Analysis of Important Part-of-Speech (PoS)

Identifying N-Grams plays a vital role in the generation of the summary. If all possible PoS tags for N (from 1 to 4) are considered, then there are approximately 1.7 million combinations of PoS tags. It will not only overload the database but also unwanted and less significant N-Grams will be considered later during the summary generation stage.

To understand which N-Grams are of greater significance, the research has analyzed a few JSON documents. The analysis starts by obtaining the sentence from the abstract and the body of articles. The period or semicolon are considered to be the line separators. An average article in CORD-19 dataset contains around 80-100 sentences.

Taking the analysis of 2-Grams as example, Figure 1 shows the Bar chart of PoS tags for 2-Grams in a chosen article.

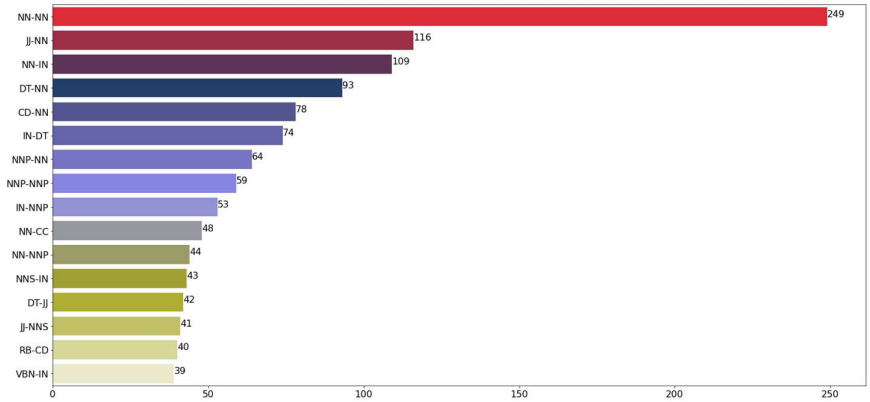

Fig. 1. The frequencies of the 2-Gram Parts of Speech in an article.

This chosen article has a high chance to see Noun-Noun followed by Adjective-Noun for 2-Gram PoS tags. Furthermore, Figure 2 shows the trends of 2-Gram PoS tags are similar in two chosen articles with line plot.

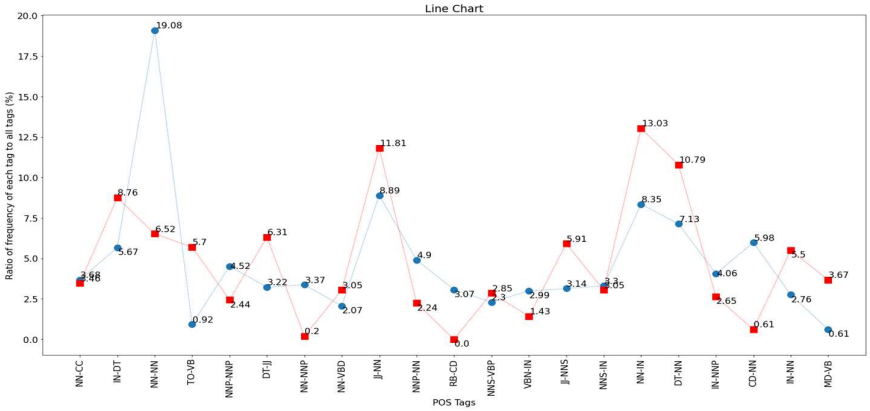

Fig. 2. The the ratio of the frequency of each tag to the frequency of all 2-Gram tags.

4 3-Stage Summary Generation

4.1 Automatic Extraction and Verification

The CORD-19 dataset is publicly available in compressed file formats. Automatic extraction and verification of the dataset involve two tasks. The first task is responsible to automatically uncompressing the compressed files uploaded to the server. The files extracted from CORD-19 dataset may contain JavaScript Object Notation (JSON) files, text files, word documents, PDFs, image files, Excel sheets, and other file formats. The second task is filtering the scholarly articles available in JSON format and removing all unnecessary files. Figure 3 shows the overview of the process.

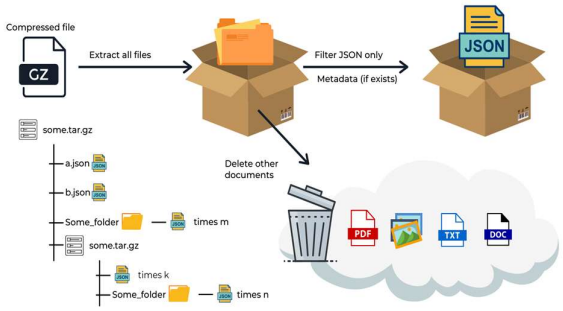

Fig. 3. The overview of Automatic Extraction and Verification of the scholarly articles.

4.2 Processing Documents

This stage is to extract all of the essential data from the JSON documents and store them in the database. After the process is complete, the original JSON document is deleted and marked as processed. Figure 4 shows the overview of the process.

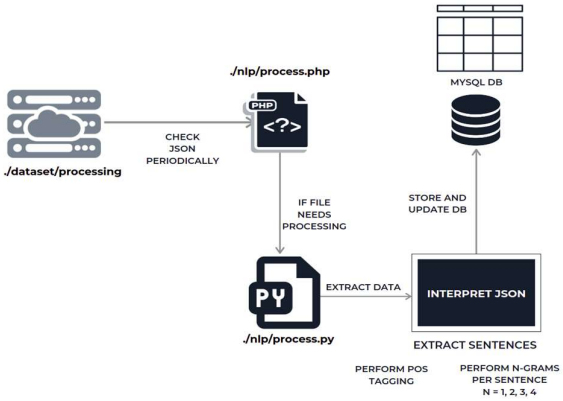

Fig. 4. The overview of Processing Documents by NLP Techniques.

The most frequent Parts of Speech (PoS) tags had been found via the analysis of few JSON documents. This PoS list is used for verifying the eligibility of the N-Grams. Next, all the sentences are extracted followed by the N-Grams, where N is from 1 to 4. The frequencies of N-Grams per document and per sentence are also counted and saved. The N-Grams act as the keyword for the article's content and user's question, and it plays a significant role in the summary generation later.

4.3 Summary Generation

Document and sentence selections play a fundamental role in generating the summary for a user's question. When a user asks a question, the first task for the system is to identify the keywords from the question. To recognize the keywords, the same validation process for identify eligible N-Grams with the most frequent PoS tag list is taken. The system ends up getting a N-Gram vector that consists of a list of eligible N-Grams extracted from the question.

The next task is to create a target vector that consists of the frequency of the extracted eligible N-Grams. Since the frequency of each N-Gram has been stored in the database, the target vectors can be generated easily. After generating the target vector, the frequency of each N-Gram per document and per sentence stored in the database is used to generate the correspondent document vectors and sentence vectors as Figure 5 shows

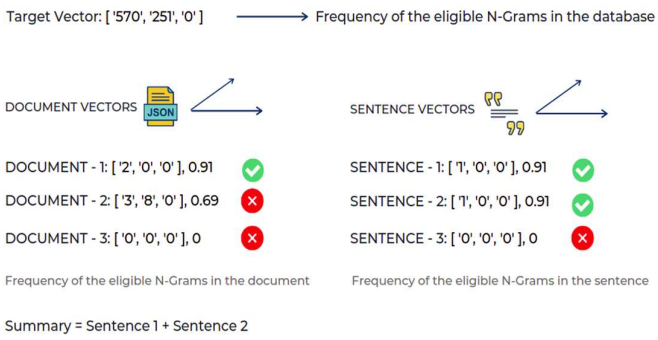

Fig. 5. Target, Document and Sentence vectors.

The cosine similarity is adopted to select the article whose document vector that matches most to the target vector. The cosine similarity score lies between -1 and +1, where +1 means two vectors are identical and -1 means they are opposite. Once the article with the highest score is found, the sentence vectors of that article are generated based on the stored frequencies of the N-Grams per sentence. Again, the cosine similarity is used to select the sentence vectors that match most to the target vector.

Figure 5 shows an example where the document vector (Document-1) with cosine similarity 0.91 has the highest score towards the target vector, so the article is chosen. Then two sentences in the article, Sentences S1 and S2, that have the best match with the target vector are also identified. Thus, the summary could be a concatenation of these two chosen sentences – the simplest way.

5 The System and The Evaluation Plan

5.1 The System

The users can access the system[1] with their browsers (see Figure 6) to ask a question related to Coronavirus as well as COVID-19. When a user asks a question, the question along with an auto-generated universally unique identifier (UUID) will be sent to the system for generating a summary. The user can then see the summary and provide his or her feedback include the perceived relevance and satisfaction scores toward the generated summary. If the user is satisfied with the summary, he or she can provide higher scores on his or her perceived satisfaction and relevance. On the other hand, if the generated summary is not good enough, then he or she can give lower scores on the perceived relevance and satisfaction. The feedback will be used to improve the summary generation method in future research.

[1] https://askcovidq.vipresearch.ca/ui/

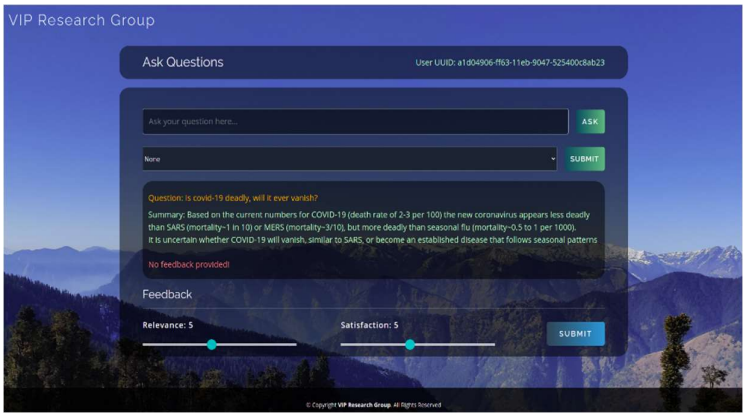

Fig. 6. The user interface for users asking question and providing their perceived relevance and satisfaction toward the generated summary.

5.2 Evaluation Plan

This research plans to use the snowball sampling method. The research team plans to invite participants starting from the supervisor's colleagues and extending the invitation to their networks. A link will be included in the invitation email/posting to the automatic summary generation service website to participate in the research.

When they access the system for the first time, a UUID will be generated as the study ID. The website will record and connect the UUID to the questions and the generated summaries as well as participants' perceived relevance and satisfaction toward the generated summaries. After they ask the system questions for three times, they will be asked to spend less than 2 minutes to fill out the System Usability Scale (SUS) which has 10 5-point Likert Scale questions. If they accept to fill out the SUS and submit their responses, the system will also record those data under the study ID.

Before starting the data analysis and evaluation process, all questions will be reviewed first to exclude those non-Coronavirus questions and the correspondent feedback and SUS responses to avoid malicious entries. The system's usability score will be calculated based on participants' SUS responses and see if the score is higher than 68 – which means the usability of the system is OK or above.

The perceived relevance and satisfaction toward the generated summaries will be assessed with descriptive statistics. For those summaries that have low perceived relevance and satisfaction ratings, the research team will further investigate the potential causes and try to adjust and fix the summary generation method accordingly. The correlation between summaries' relevance and satisfaction ratings will also be tested. The expectation is to see a very high correlation. That means when a summary is highly related to the question participants ask, they would be more satisfied with the summary.

6 Conclusion

The research proposed in this paper is an introductory approach for generating a summary based on the user's question. In particular, this paper discusses the methods involved in extracting and processing useful information from the large corpus of scholarly articles (i.e., CORD-19 dataset) with NLP techniques. In addition, the paper also explains how to store the essential information in the database to use it in the summary generation. Finally, with the Cosine Similarity, the comparison results for the generated document and sentence vectors to the target vector could be used for generating the summary for the asked question.

The research has following limitations. First, only a simple analysis on identifying the important PoS tags has been done. Although some reasonable results are found, it does not guarantee that the PoS tag list we consider eligible will be always true. The list can change at any time depending on the frequency of the N-grams seen by the computer and can be different for articles in different application domains and/or research areas. Second, the time cost for the proposed system to generate a summary for a given question is strongly influenced by the number of articles that the system has processed and the number of keywords in the question. When the system reads and processes more articles, then the current method will need more time to identify document(s) and its sentences for generating a summary paragraph.

Instead of having the system a built-in and hardcoded PoS tag list, the system could access a service[2] that learns, calculates, and updates the PoS tag list based on DBpedia frequently. An optimized algorithm that can significantly reduce the time to generate the summary will be further investigated. Furthermore, the collected feedback will be used to find the loopholes of the system and provide a better and more relevant summary.

References

1. Singh, S.: Natural Language Processing for Information Extraction. Ithaca, New York: arXiv. https://arxiv.org/abs/1807.02383 (2018)
2. Hamid Mughal, M. J.: Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview. International Journal of Advanced Computer Science and Applications, 9(6), 208-215. http://dx.doi.org/10.14569/IJACSA.2018.090630 (2018)
3. Mana, T., Zhukovab, N.A., Thawa, A.M., Abbas, S.A.: A decision support system for DM algorithm selection based on module extraction. Procedia Computer Science, 186, 529-537. https://doi.org/10.1016/j.procs.2021.04.173 (2021)
4. Jain, A., Kulkarni, G., Shah, V.: Natural Language Processing. International Journal of Computer Sciences and Engineering, 6(1), 161-167. https://doi.org/10.26438/ijcse/v6i1.161167 (2018)
5. Romanov, A., Lomotin, K., Kozlova, E.: Application of Natural Language Processing Algorithms to the Task of Automatic Classification of Russian Scientific Texts. Data Science

[2] https://ws-nlp.vipresearch.ca/ngram_pos

Journal, 18, Article 18:37. London: Ubiquity Press. https://datascience.codata.org/articles/10.5334/dsj-2019-037/ (2019)
6. Bahja, M.: Natural Language Processing Applications in Business. In: Wu, R. M. X., Mircea, M. (eds.) E-Business - Higher Education and Intelligence Applications. London: IntechOpen. https://doi.org/10.5772/intechopen.92203 (2021)
7. Kowalski, R., Esteve, M., Mikhaylov, S. J.: Improving public services by mining citizen feedback: An application of natural language processing. Public Administration, 98(4), 1011-1026. https://doi.org/10.1111/padm.12656 (2020)
8. Fu, Y., Feng, Y., Cunningham, J. P.: Paraphrase Generation with Latent Bag of Words. Ithaca, New York: arXiv. https://arxiv.org/abs/2001.01941 (2020)
9. Elton, D. C., Turakhia, D., Reddy, N., Boukouvalas, Z., Fuge, M. D., Doherty, R. M., Chung, P. Q.: Using natural language processing techniques to extract information on the properties and functionalities of energetic materials from large text corpora. Ithaca, New York: arXiv. https://arxiv.org/abs/1903.00415 (2019)
10. Qader, W. A., Ameen, M. M., Ahmed, B. I.: An Overview of Bag of Words:Importance, Implementation, Applications, and Challenges. In Proceedings of 2019 International Engineering Conference (IEC), Erbil, Iraq, June 23-25, 2019. https://doi.org/10.1109/IEC47844.2019.8950616 (2019)
11. Mullen, L. A., Benoit, K., Keyes, O., Selivanov, D., Arnold, J.: Fast, Consistent Tokenization of Natural Language Text. Journal of Open Source Software, 3(23), 655, https://doi.org/10.21105/joss.00655 (2018)
12. Ding, C., Utiyama, M., Sumita, E.: NOVA: A Feasible and Flexible Annotation System for Joint Tokenization and Part-of-Speech Tagging. ACM Transactions on Asian Low-Resource Language Information Processing, 18(2), Article 17. https://doi.org/10.1145/3276773 (2018)
13. Al-Radhi, M. S., Csapó, T. G., Németh, G.: Advances in Speech Vocoding for Text-to-Speech with Continuous Parameters. Ithaca, New York: arXiv. https://arxiv.org/abs/2106.10481 (2021)
14. Kaur, S., Agrawal, R.: A Detailed Analysis of Core NLP for Information Extraction. International Journal of Machine Learning and Networked Collaborative Engineering, 1(1), 33-47. https://ssrn.com/abstract=3376818 (2018)
15. Zhong, M., Liu, P., Chen, Y., Wang, D., Qiu, X., Huang, X.: Extractive Summarization as Text Matching. In Proceedings of Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 5-10, 2020. 6197–6208. http://dx.doi.org/10.18653/v1/2020.acl-main.552 (2020)
16. Gunawan, D., Sembiring, C. A., Budiman, M. A.: The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents. Journal of Physics: Conference Series, 978, Article 012120. http://dx.doi.org/10.1088/1742-6596/978/1/012120 (2018)
17. Papagiannopoulou, E., Tsoumakas, G.: Local Word Vectors Guiding Keyphrase Extraction. Ithaca, New York: arXiv. https://arxiv.org/abs/1710.07503 (2018)
18. Shahmirzadi, O., Lugowski, A., Younge, K.: Text Similarity in Vector Space Models: A Comparative Study. Ithaca, New York: arXiv. https://arxiv.org/abs/1810.00664 (2018)
19. Farouk, M.: Measuring Sentences Similarity: A Survey. Ithaca, New York: arXiv. https://arxiv.org/abs/1910.03940 (2019)
20. Alam, F., Afzal, M., Malik, K. M.: Comparative Analysis of Semantic Similarity Techniques for Medical Text. In Proceedings of 2020 International Conference on Information Networking (ICOIN), Barcelona, Spain, January 7-10, 2020, 106-109. https://doi.org/10.1109/ICOIN48656.2020.9016574 (2020)