WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Text Summarization using NLP and Deep Learning Techniques

Evangelin Sonia[1], Brindha Shanmugavadivel[2]

[1]Assistant Professor, Dept. of Computer Science Engineering, Karunya Institute of Technology & Sciences, Coimbatore, Tamil Nādu
[2] Dept. of Computer Science Engineering, Sri Shakthi Institute of Engineering & Technology, Coimbatore, Tamil Nādu

ABSTRACT –

Text Summarization is defined as a vital task in Natural Language Processing that extracts a brief summary from a huge dataset. With a wide range of data available across the internet, it is very difficult to read the complete dataset and understand the context. Hence, inorder to quickly gather the crisp information from a huge dataset, text summarization tools came into existence. Text Summarization tools could simply collect the important information from a large data without affecting the actual context and vital information in it. Summarization techniques can broadly be categorized as Extractive and Abstractive. This paper uses the NLP techniques to semantically analyze the dataset and understand the grammatical structure of the dataset using pre-trained SpaCy NLP pipeline.

Key Words: text summarization, natural language processing, spacy, sentence score, extractive summarization.

1. INTRODUCTION

Text Summarization is the process of extracting a brief summary from a lengthy text document without affecting the actual meaning of the dataset. It has become very much important for us due to our busy daily routine. Everyone prefers to understand an article or a news in a succinct way rather than reading a lengthy document and understanding the context by ourselves which is tedious and time consuming. The objective is to only remove the text data that doesn't change the meaning of the content.

Fig -1.1: Text Summarization overview

1.1 Types of Text Summarization

There are no fixed methods to categorize the summarization process but based on the usage or purpose of summarization, we classify them into following types as below:

- Short tail, long tail, Single/Multiple entity, General purpose, Domain specific, Informative, Headline generation, Keyword extraction based summarization methods.

However, the final text summarization process is broadly classified as two types namely:

- **Extractive Summarization**, which gathers the important phrases from the actual dataset and collates it together to generate a short simple summary.
- **Abstractive Summarization** involves more of NLP tasks by adding new terms and words that doesn't affect the actual context

Basic text summarization steps involve:

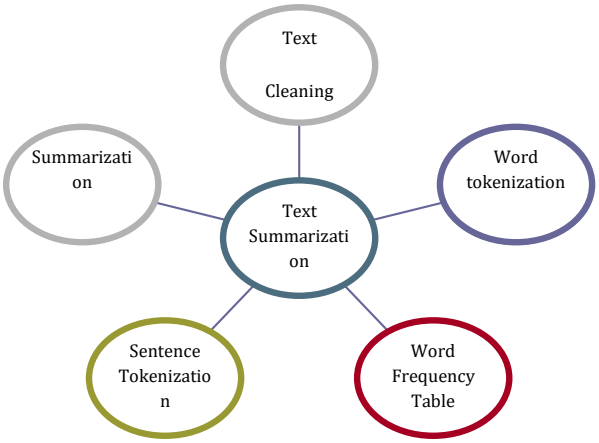

Fig -1.2: Text Summarization Steps

1.2 Strategies to identify key sentences for Text Summarization

To gather important sentences for the concise summary, a list of features as below, can be considered:

- Term Definition, Location, Title/Headline Word, Cue Method, Sentence Length, Similarity, Proximity, Proper noun.

1.3 Applications of Text Summarization

Kharade [5] et al., discussed about few applications of Text Summarization tools that include:

- ✓ Newsletters
- ✓ Scientific Research
- ✓ Social Media Marketing
- ✓ Content Writing
- ✓ Financial Research
- ✓ Video Scripting, etc.,

1.4 Need for Text Summarization

Text Summarization can be beneficial in many ways like:

- ✓ Optimize reading time
- ✓ Simplify report investigation process
- ✓ Improved business content archival process
- ✓ Improved question-answer based system

2. Problem Definition

With an exponential increase in dataset across the internet world, it is tedious to read through and understand the exact context of a large document base in quick span of time. Hence, a text summarization tool has been proposed to optimize the reading time of common people.

3. RESEARCH METHODOLOGY AND RELATED WORK

N. Moratanch[1] et al., Automatic text summarization is the process of producing a short and meaningful summary while preserving key information content of the original document. Pratibha Devi Hosur[2] et al., suggested a system by implementing unsupervised learning during automatic text summarization in which the overall depiction used NLP tasks which included input text document, text pre-processing, lesk algorithm and finally, generating the concise summary.

G. Vijay Kumar[3] et al., discussed about the regular patterns about both abstractive and extractive techniques that are used for summarization of text. Gogulamudi[4] et al., discussed about the text rank-based approach for text summarization which is a diagram-based positioning model for long text processing that is used to find the most relatable sentences in dataset and also to find keywords. Text rank algorithm is similar to page rank algorithm which uses sentences instead of web pages. Munot[6] et al., discussed about the comparative study of various approaches used for text summarization. Below table [Table 3.1][7][8][9] shows a study on few of the above discussed approaches.

Table -3.1: Comparative Study of Text Summarization Techniques

Techniques	Description	Content Selection	Summary Generation	Merits	Demerits
Tree Based	Based on dependency tree	theme intersection algorithm	surge language generator	Simple to implement	Does not guarantee a logical summary
Template Based	User can create a template of what should be in the summary	linguistic patterns or extraction rules	ie based md summarization algorithm	Depends on end user's requirement for summary generation	Template creation is tedious
Ontology Based	Technique is used for creating ontology	classifier	news agent	Utilizes data pre-processing, semantic info, extraction & ontology generation	only domain experts can construct ontology which is tedious
Semantic graph based	Creates a rich sematic graph (RSG) on the source text, condensing it, and then creates a final concise summary	Syntactical and morphological tags	Condensed RSG using domain ontology techniques	It generates crisp, logical, grammatically correct, and less repetitive summary	Suitable for single document summarization only
Lead and Body Phrase based	Depending on stages(insert and replace) important sentences to overwrite the leading sentence	maximum phrases of same head in lead and body sentences	insertion and substitution operations on phrases	Used for sematic based summary	Possibility of having grammatical errors

Inorder to summarize huge datasets, we have studied and analyzed many approaches as discussed in Table 3.1. Of all the methods studied, Text Rank algorithm[10][11] best fits the scenario in terms of implementation, time consumption to summarize text, logical order of sentences, grammatical correctness of sentences, etc.,

4. IMPLEMENTATION

4.1 Techniques Used

This paper focusses on the usage of SpaCy language model and python programming with a Text rank algorithm to implement the text summarization tool. SpaCy, a pre-trained NLP pipeline is used here inorder to understand the grammatical structure of the dataset. While NLTK library model provides access to many algorithms to get summarization process done, SpaCy provides the best way to do it. It provides the fastest, logical and most accurate syntactic analysis of any NLP library released to date. It also offers access to larger word vectors that are easier to customize.

4.2 Sample Coding and Output Results

```
def generate_summary(text_without_removing_dot, cleaned_text):
    sample_text = text_without_removing_dot
    doc = nlp(sample_text)
    sentence_list=[]
    for idx, sentence in enumerate(doc.sents): # we are using spacy for sentence t
        sentence_list.append(re.sub(r'[^\w\s]','',str(sentence)))

    stopwords = nltk.corpus.stopwords.words('english')

    word_frequencies = {}
    for word in nltk.word_tokenize(cleaned_text):
        if word not in stopwords:
            if word not in word_frequencies.keys():
                word_frequencies[word] = 1
            else:
                word_frequencies[word] += 1

    maximum_frequncy = max(word_frequencies.values())

    for word in word_frequencies.keys():
        word_frequencies[word] = (word_frequencies[word]/maximum_frequncy)


    sentence_scores = {}
    for sent in sentence_list:
        for word in nltk.word_tokenize(sent.lower()):
            if word in word_frequencies.keys():
                if len(sent.split(' ')) < 30:
                    if sent not in sentence_scores.keys():
                        sentence_scores[sent] = word_frequencies[word]
                    else:
                        sentence_scores[sent] += word_frequencies[word]


    summary_sentences = heapq.nlargest(7, sentence_scores, key=sentence_scores.get
```

Fig -4.1: Text Summarization Function

```
'\n\nThe researchers say a possible explanation for this warming bias may lie in a "multiplier effect," whereby a mod
est degree of warming -- for instance from volcanoes releasing carbon dioxide into the atmosphere -- naturally speeds
up certain biological and chemical processes that enhance these fluctuations, leading, on average, to still more warm
ing." In other words, there were far more warming events -- periods of prolonged global warming, lasting thousands to
tens of thousands of years -- than cooling events.'
```

Fig -4.2: Summarized Output

5. CONCLUSIONS

Hence text summarization process has been implemented with help of Text Rank algorithm and a pretrained SpaCy pipeline library model. The summarized text is observed to be more logical and grammatically correct that has much helped the readers to optimize their reading time and aptly understand the context of the dataset. As an enhancement to the existing system, dataset and algorithms can be further explored to retain the case sensitivity of the dataset, to revisit the logical order of the summarized text and to normalize long sentences in the dataset.

References

[1]. D. Kornack and P. Rakic, “Cell Proliferation without Neurogenesis in Adult Primate Neocortex,” Science, vol. 294, Dec. 2001, pp. 2127-2130, doi:10.1126/science.1065467.

[2]. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[3]. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[4]. K. Elissa, “Title of paper if known,” unpublished.

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

[6]. N. Moratanch, S.Chitrakala, Anna University, CEG, Chennai, IEEE International Conference on Computer, Communication, and Signal Processing (ICCCSP-2017)

[7]. Gogulamudi, Vijay & Yadav, Arvind & Vishnupriya, B. & Lahari, M. & Smriti, J. & Reddy, D. (2021). Text Summarizing Using NLP. 10.3233/APC210179.

[8]. Kharade, Kabir & Katkar, Smita & Patil, N. & Sonawane, V. & Kharade, Shraddha & Pawar, T. & Kamat, Rajanish. (2021). Text Summarization of an Article Extracted from Wikipedia Using NLTK Library. 10.1007/978-3-030-88244-0_19.

[9]. Munot, Nikita & Govilkar, Sharvari. (2014). Comparative Study of Text Summarization Methods. International Journal of Computer Applications. 102. 33-37. 10.5120/17870-8810.

[10]. F. Chen, K. Han, and G. Chen, "An approach to sentence-selection based text summarization," in TENCON'02. Proceedings. 2002 IEEE Region 10Conference on Computers, Communications, Control and Power Engineering, vol. 1. IEEE, 2002, pp. 489-493.

[11]. Text Summarization: An Essential Study [Prabhudas Janjanam and CH Pradeep Reddy