

Text Summarizer and Translator using NLP

Mrs A. M. Chimann¹, Mr Yashraj Patil² and Mrs M.S. Dabade³

^{1,3}Asst.Professor Walchand College of Engineering, Sangli.

² Student, Walchand College of Engineering, Sangli.

Email: amruta.umrani@walchandsangli.ac.in, yashrajpatil@walchandsangli.ac.in, manisha.dabade@walchandsangli.ac.in

Abstract— In this new era, where tremendous information is available on the Internet, it is most important to provide an improved mechanism to extract the information quickly and most efficiently. It is very difficult for human beings to manually extract the summary of a large document of text. There is plenty of text material available on the Internet. So there is a problem of searching for relevant documents from the number of documents available, and absorbing relevant information from it. In order to solve the above two problems, the automatic text summarization is very much necessary. Text summarization is the process of identifying the most important meaningful information in a document or set of related documents and compressing them into a shorter version preserving its overall meanings. The integration of translation functionality with the text summarization technique will add on to the utility of the system.

Index Terms— Tokenization, NLP, Sentence scoring

I. INTRODUCTION

Extracting meaningful information from internet material is critical in today's information-rich world. Manual summary of huge papers takes time, and obtaining relevant information from many sources is difficult. Automatic text summarization has emerged as a crucial answer, condensing critical information while maintaining meaning. For efficient multilingual information extraction, our project integrates Spacy and the My Memory Translation API. This combo improves text summarization while also providing translation functionality. Related research investigates algorithms and approaches for text summarization and machine translation, such as statistical analysis, machine learning, and deep learning. These innovations have increased the precision and fluency of the summarization and translation processes.

A. Problem statement

Design and Implementation of a Text Summarization and Translation system using NLP.

B. Objectives

1. Study and implementation of Spacy.
2. To implement the text summarization module.
3. To implement the translation module.
4. To integrate the model with a web application.

II. METHODOLOGY

1. The text is taken as input.

2. The process of text cleaning is done by removing all the punctuations.
3. Then, text formatting is done.
4. Before starting the tokenization process, we load all the necessary modules and create an object referred to as NLP for the process and model.
5. In the tokenization process, the sentences and word and model it.
6. The process of scoring and selection is done.
7. Finally, we can translate the obtained summary into any target language.

A. Spacy

Spacy is an open-source natural language processing (NLP) package with sophisticated and efficient text processing features. Tokenization, part-of-speech tagging, named entity recognition, and dependency parsing constitute some of its many features. Spacy enables developers to process text data in a variety of linguistic contexts by supporting several languages and pre-trained models. It is well-known for its speed and performance, as well as its ability to customize and train models for specific domains or languages. Spacy's comprehensive capabilities and user-friendly interface have acquired appeal among researchers and developers, making it a significant tool in the NLP field. "en_core_web_sm" refers to one of the pre-trained English models provided by Spacy. It is a small-sized model that includes vocabulary, syntax, and entities trained on various English text sources.

B. My Memory

My Memory is a well-known translation memory and machine translation platform with a large database and a diverse set of language pairs. It provides easy web-based access as well as an API for smooth integration into translation operations. My Memory is a wonderful resource for precise and efficient language translation because to its strong translation services.

III. GENERAL FLOW OF THE UNDERLYING MODEL

A. Text Cleaning

Text cleaning is an essential step in automatic summarization. It involves removing punctuation, stop words, lowercasing, and eliminating special characters like emojis, hashtags, and URLs. Additionally, techniques like stemming and lemmatization help reduce words to their simplest form.

B. Text Formatting

Text formatting is critical in the automatic text summarization process. To produce meaningful and accurate summaries, the summarization algorithms rely on consistent and structured input text.

C. Tokenization

Tokenization is an important step in the automatic text summarization process that involves dividing the input text into smaller units known as tokens.

D. Sentence Scoring

Scoring is an important part of the automatic text summarization process because it allows the algorithm to recognize and prioritize the most important information in the input text.



Figure 1. Sentence Scoring

E. Selection

Selection is a crucial phase in the automatic text summarization process since it includes deciding which sentences or phrases to include in the summary.

F. Translation

The summary is translated into the target language using machine translation techniques.

G. Flow Diagram

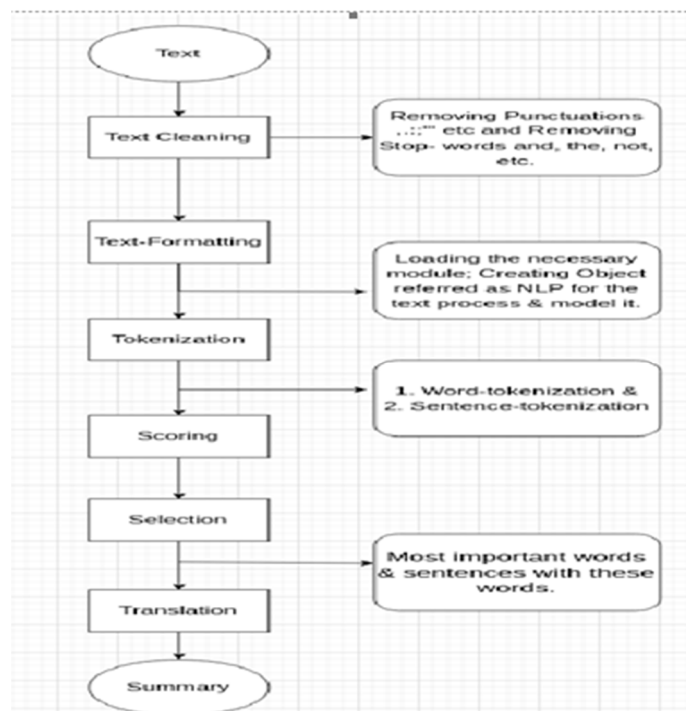


Figure 2. Flow Diagram

H. Testing (Unit, Integration and System)

For demonstration purposes, we are actively adding an input paragraph written in Hindi. It is worth noting, however, that we have the luxury of using numerous Indian regional languages or even English as our source language. By taking into account such a wide range of languages, we want to handle the complexities and nuances unique to each language, so broadening the usefulness and effectiveness of our project in multilingual environments

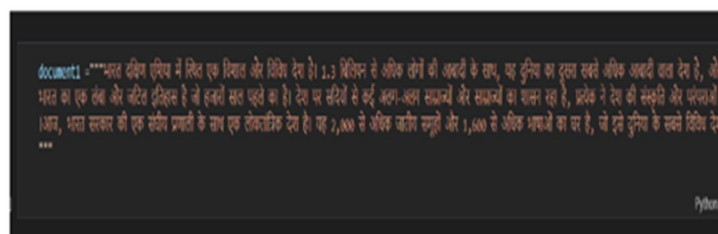


Figure 3. Input Paragraph in Source language

We not only chose Marathi as the target language for translation in our demonstration, but we additionally employed text summarization. The provided text has been summarized, resulting in a reduced version that captures the vital data. Furthermore, the summarized content has been accurately translated into Marathi. This highlights our system's integrated text summarizing and translation functionality.

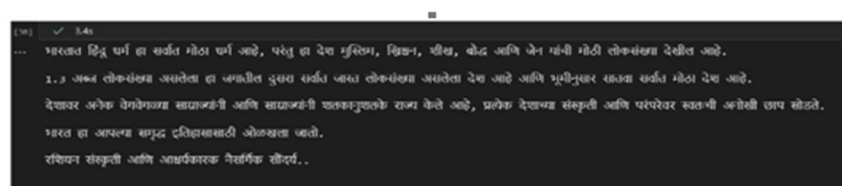


Figure 4. Output Paragraph in Target language



Figure 5. Output showing paragraph length difference

IV. RESULTS AND CONCLUSION

The Input screen has a text box for taking input and a dropdown to select the target language



Figure 6 Home Page



Figure 7. Output Page

REFERENCES

- [1] Rose, S., Engel, D., Cramer, N. and Cowley, W., 2010. Automatic keyword ex-traction from individual documents. Text mining: applications and theory, pp.1-20.
- [2] H. P. Luhn, "The Automatic Creation of Literature Abstracts," in IBM Journal of Research and Development, vol. 2, no. 2, pp. 159-165, Apr. 1958, doi: 10.1147/rd.22.0159.
- [3] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Text. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.
- [4] Steven Bird and Edward Loper. 2004. NLTK: The Natural Language Toolkit. In Proceedings of the ACL Interactive Poster and Demonstration Sessions, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.
- [5] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and
- [6] David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Lin-guistics: System Demonstrations, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.