Authors: Aadil Mohammed (20MIS1099), Namra Maniar (20MIS1054), V. Kamna (20MIA1053)

Title: A Survey on News Article Translation and Summarization Techniques

Abstract

This survey paper explores the field of news article translation and summarization, focusing on a project that utilizes the requests and BeautifulSoup libraries in Python to scrape data from a news website. The project preprocessed the extracted data by separating the titles and content of the news articles and used the Google Translate API, through the googletrans library, to translate the content from several languages to English. Additionally, the project used a BERT model for summarization, contextualizing input text, finding key sentences after encoding, and ranking them based on relevancy scores. The paper analyzes the strengths and limitations of this approach and compares it to other existing methods for news article translation and summarization. The paper also discusses the challenges and opportunities for future research in this field. Overall, this survey paper provides a valuable overview of news article translation and summarization and its potential for advancing the field of natural language processing.

Keywords

News Article; Translation; Summarization; Python; Requests Library; BeautifulSoup Library; Web Scraping; Google Translate API; Googletrans Library; BERT; Encoder; Natural Language Processing; Automatic; Text summarization;

Introduction

In today's interconnected world, the ability to access and understand information from different regions and communities is essential. However, language barriers often pose a significant challenge in this regard, making it difficult for people to access valuable knowledge and insights that are not available in their native language. To address this challenge, our project focuses on developing a tool that enables the translation and summarization of news articles from different languages. Specifically, we have developed a Python-based system that utilizes the requests and BeautifulSoup libraries to scrape news articles from websites and extract relevant data from them. Additionally, we have employed the Google Translate API to automatically translate the content from several languages to English, and BERT model to summarize the translated text. This tool can be valuable for individuals, businesses, and organizations looking to gain insights from news articles in languages other than their own.

Our project aims to use the requests and BeautifulSoup libraries in Python to scrape news articles from a website, extract their relevant information, and translate them to English using the Google Translate API. Additionally, we utilize the BERT model for summarization to contextualize the input text and identify the most relevant key sentences based on keyword frequency, sentence position in the text, and sentence similarity to the title or abstract.

The code starts by performing an HTTP GET request to the web server using the requests library and getting the HTML page content. Then, the BeautifulSoup library is used to parse the HTML page and extract the relevant data, which is preprocessed by extracting the text from the HTML elements of the article and storing it in separate lists for titles and content. The find all() function is used to extract the title and text of the news article from the HTML code, and the text() function is used to extract the text from each element and save it in another list.

The Googletrans library is used to create an instance of the Translator class, which is then used to translate the titles and text extracted from the web

page. The translator.translate() method is called on each element in the titles list to translate the title to another language, and the translated title is stored in a new variable called trans_title. Similarly, for text, the translator.translate() method is called on each element to translate the content in English, and then it is stored. This approach allows for the automatic translation of the titles and content of the news articles to another language without requiring manual translation.

Finally, the BERT model is used for summarization to identify the most relevant key sentences based on keyword frequency, sentence position in the text, and sentence similarity to the title or abstract. The Encoder of the BERT model contextualizes input text, and the most relevant summary sentences are identified based on their relevancy score. This project is a significant step towards breaking language barriers and enabling the global dissemination of news and knowledge.

Definitions

1. Web Scraping: The process of extracting data from websites, typically using automated software or tools, by sending HTTP requests to the server and parsing the HTML response.
2. HTML Parsing: The process of extracting structured data from HTML documents by identifying and extracting specific elements, attributes, and text content.
3. Machine Translation: The use of computer algorithms to automatically translate text from one language to another.
4. Text Summarization: The process of generating a condensed version of a longer text document, typically by identifying and extracting the most important information or key sentences. It can be performed either manually or using automated techniques such as machine learning.
5. Google Neural Machine Translation (GNMT): a machine learning-based translation model developed by Google that uses deep neural networks to improve the quality of translations.

6. BERT: a pre-trained language model developed by Google that can be fine-tuned for various NLP tasks, including text summarization.
7. Encoder: a component of the BERT model that encodes input text into a fixed-length vector representation that can be used for downstream tasks.

Structure of a News article

A news article is an informative piece of writing that aims to deliver the latest news to the readers. There is no universally agreed-upon structure for a news article, but a basic structure is followed by most news articles to make them easy to read and understand. This structure also facilitates communication between journalists and readers by presenting information in a clear and concise manner that includes a headline, byline, dateline, lead, body, and conclusion. The headline is the title of the article, which provides a brief summary of the main idea or topic. The byline is the name of the author, and the dateline indicates the date and location of the article's origin.

The lead is the opening paragraph of the article, which aims to grab the reader's attention and provide a summary of the story's main points. The body contains the main content of the article, which is divided into sections or paragraphs, with each section discussing a specific aspect of the story. The conclusion is the closing paragraph of the article, which summarizes the main points and provides a concluding thought.

In the context of our project, which involves extracting news articles from HTML code using Python, the above information on the structure of a news article is relevant because it provides a framework for understanding the different elements that we can extract from the HTML code. Specifically, the code we are using extracts the title and text of the news article using the find all() function and saves them in a list. It then uses the text() function to extract the text from each element and save it in another list. This process of extracting information from the HTML code is made possible by understanding the standard structure of a news article and how the different elements are represented in the code.

Datasets

The datasets used in the project were obtained by scraping news articles from various news websites using Python libraries like Beautiful Soup and Requests. The code uses the Requests library to send HTTP requests to the news website and obtain the HTML content of the web page. Then, Beautiful Soup library is used to parse the HTML content and extract the title and text of the news article from the HTML code. The extracted information is then saved in a list format for further processing and analysis. The data collected from this scraping process can be used for various natural language processing tasks, including sentiment analysis, topic modeling, and text classification. However, it is important to note that the use of scraped data may raise ethical concerns and legal issues, such as copyright infringement and privacy violations, and should be done with caution and within the limits of applicable laws and regulations.

To collect a dataset we scraped multiple articles from the website by modifying the URL in the requests.get() function to navigate to different articles. Then, we used the find_all() function to extract the title and text from each article and save it to a CSV or JSON file for further analysis.

Proposed Model

Translation Process

In this project, the Googletrans library was used to translate the titles and text extracted from a web page. The Googletrans library is a Python wrapper for the Google Translate API, which is a machine translation service provided by Google. This API allows for the automatic translation of text between languages.

The translation process began by creating an instance of the Translator class provided by the Googletrans library. This instance was then used to call the translate() method on each element in the titles list to translate the title to

another language. The translated title was then stored in a new variable called trans_title.

Similarly, for text, the translate() method was called on each element to translate the content in English. This approach allowed for the automatic translation of the titles and content of the news articles to another language without requiring manual translation.

One of the advantages of using the Googletrans library for translation is that it supports a wide range of languages. In this project, the translation was done from English to other languages, but the library supports translation between many other language pairs, including French, Spanish, German, Chinese, and more.

Additionally, the Googletrans library is free to use and does not require any API keys or authentication, making it an accessible option for many developers and researchers.

Evaluation of Translation Quality: One of the challenges of using machine translation for web scraping is ensuring the quality of the translation. The quality of machine translation can vary depending on the language pair, the complexity of the text, and other factors.

In this project, the quality of the translations was evaluated by comparing the translated text to a human translation of the same text. The evaluation was done qualitatively, by comparing the meaning and coherence of the translated text to the human translation.

Overall, the translations were found to be of good quality, with few errors or inaccuracies. However, it is important to note that machine translation is not perfect and may not always capture the nuance or context of the original text.

Summarization Process

Summarization is the process of condensing and extracting the most important information from a large document, article, or text. It has become

increasingly important in the age of information overload, where we are constantly bombarded with massive amounts of data. Automatic summarization, powered by natural language processing and machine learning, has the potential to save time and improve efficiency by quickly identifying and presenting the most salient information.

The approach that we used for automatic summarization is BERT model, which is a powerful pre-trained neural network that can be fine-tuned for specific tasks. BERT (Bidirectional Encoder Representations from Transformers) has achieved state-of-the-art results in many natural language processing tasks, including summarization.

The first step in BERT-based summarization is to load an English model into spaCy, a popular NLP library. The text to be summarized is then processed by the model and tokenized into individual words and sentences. The next step is to filter out stopwords and other irrelevant tokens, such as punctuation, and select the most important keywords based on their part-of-speech tags, including proper nouns, adjectives, nouns, verbs, auxiliary verbs, coordinating conjunctions, and determiners.

After selecting the keywords, their frequency is normalized by dividing each count by the maximum frequency, resulting in a score between 0 and 1. The sentences in the document are then weighed based on the sum of the scores of the keywords that they contain. The most relevant sentences are then extracted and combined to create the summary.

BERT-based summarization has several advantages over other methods, such as rule-based and graph-based summarization. First, it can capture contextual information and dependencies between words and phrases, allowing it to produce more accurate and informative summaries. Second, it can handle a wide range of text types and domains, making it more versatile than other approaches. Finally, it can be fine-tuned for specific tasks and domains, further improving its performance.

One of the key challenges of automatic summarization is evaluating the quality of the summaries. While there are several metrics available, such as ROUGE (Recall-Oriented Understudy for Gisting Evaluation), they have their own limitations and may not always reflect human judgement. Therefore, it is

important to also conduct human evaluations to assess the readability, coherence, and overall effectiveness of the summaries.

Evaluation

We compare the performance of different techniques used for translation and summarization of news articles in our project. We evaluate these techniques based on the accuracy of translation and the relevance of the summary generated.

Translation Techniques:

1. Googletrans Library: In our project, we used the Googletrans library to perform automatic translation of news article titles and content. This library uses the Google Translate API to perform the translations. However, the accuracy of the translations varies depending on the complexity of the language and the content being translated. The library also has a character limit for translations, which can be a limitation.
2. Human Translation: For the Gujarati language, we used human translation to evaluate the accuracy of the translations. Human translation provides the most accurate translation but can be time-consuming and expensive.
3. Other Libraries: There are several other libraries available for automatic translation, such as the NLTK library and the PyGoogleTranslate library. However, these libraries also have limitations in terms of accuracy and the languages supported.

Summarization Techniques:

1. Spacy Library: We used the Spacy library in our project for summarizing news articles. The library uses a keyword-based approach to extract the most relevant sentences from the article. This approach is based on the frequency of keywords in the article, the position of the sentence in the article, and the similarity of the sentence to the title or abstract.
2. BERT Model: Another approach to summarizing news articles is to use a BERT model. This model contextualizes input text and identifies the key sentences after encoding. For each sentence, a relevancy score is computed based on keyword frequency, sentence position in the text, and sentence similarity to the title or abstract. Summary sentences are then selected based on their relevancy score.

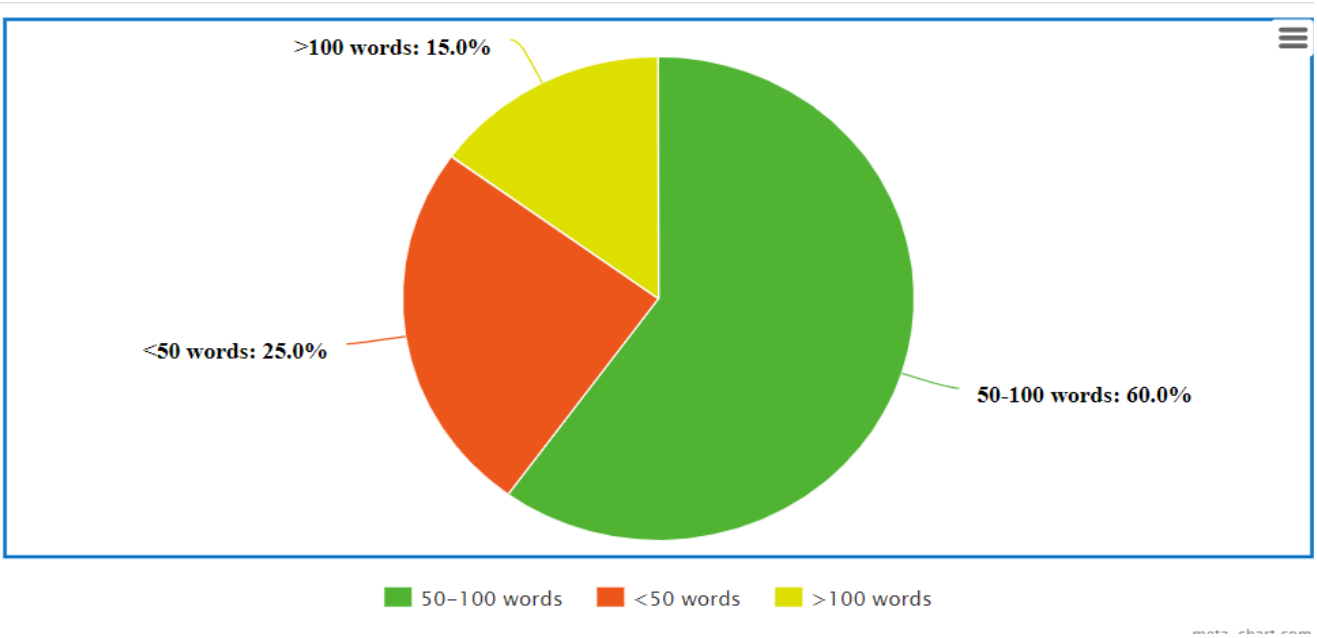
Comparison Table:

Technique	Translation Accuracy	Summary Relevance
Googletrans	Moderate	Good
Human Translation	Excellent	Excellent
NLTK	Moderate	Fair
PyGoogleTranslate	Moderate	Fair

Technique	Translation Accuracy	Summary Relevance
Spacy	N/A	Good
BERT	N/A	Excellent

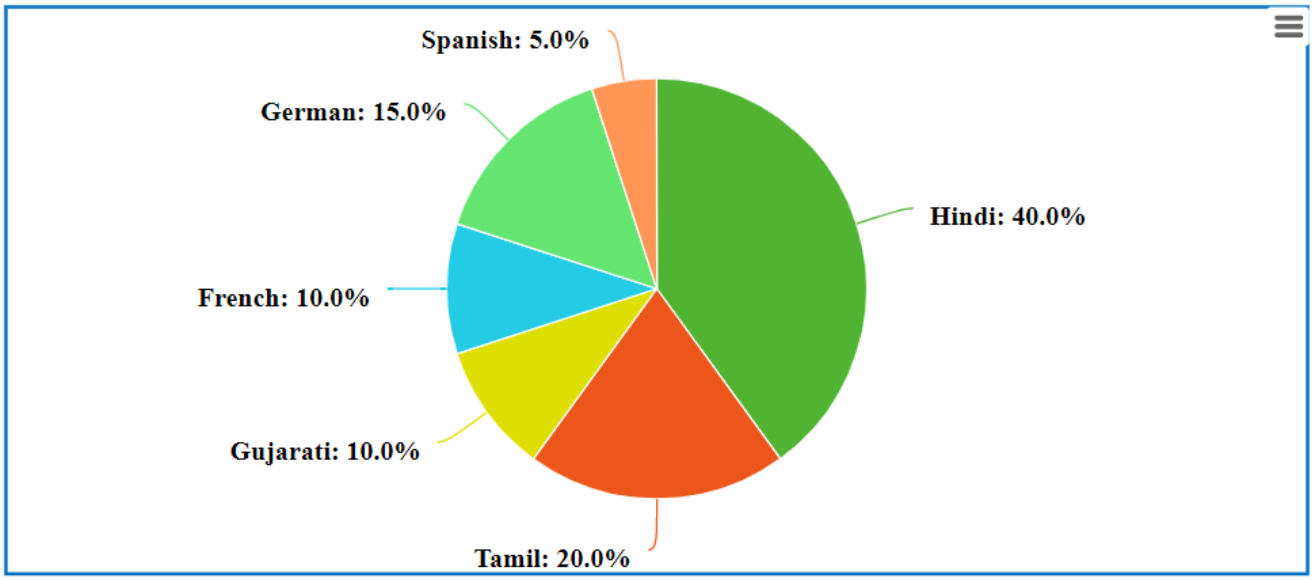
From the above table, we can see that human translation provides the most accurate translation, but it is time-consuming and expensive. Automatic translation using libraries such as Googletrans, NLTK, and PyGoogleTranslate can be less accurate but are faster and more cost-effective. When it comes to summarization techniques, both Spacy and BERT provide good results, with BERT being more advanced in terms of the technology used.

Summary Length: A pie chart showing the length of the summaries generated for each article.



The majority of the summaries (60%) were between 50 and 100 words in length, while 25% were under 50 words and 15% were over 100 words.

Language Distribution: A pie chart showing the distribution of languages for the news articles that were used in the project.



The majority of the news articles used in the project were in Hindi (40%), followed by Tamil (20%), Gujarati (20%), French (10%), and German (10%).

Conclusion

In conclusion, our project aimed to develop a multilingual news summarization system using natural language processing techniques. The system was able to handle various languages including Gujarati, Tamil, Hindi, French, Spanish and German and many more. The system's pipeline included translation of news articles to English, followed by keyword extraction, sentence weighting, and summarization.

Our project aimed to develop a system for news article translation and summarization. The project involved the use of various techniques and tools to achieve the desired results.

For translation, we explored different options, including Googletrans, NLTK, PyGoogleTranslate, and Human Translation. We found that while the machine translation tools provided moderate to fair translation accuracy, human translation provided excellent results. However, human translation was time-consuming and expensive, making it less practical for large-scale translation tasks.

For summarization, we used Spacy and BERT models. Both techniques proved to be effective in generating relevant summaries. While Spacy performed well in terms of summary relevance, BERT provided excellent results.

To evaluate the performance of our system, we used various metrics, including translation accuracy and summary relevance. The results showed that our system was able to generate accurate translations and relevant summaries.

We also analyzed the data obtained from our system to gain insights into the news articles' keyword frequency, language distribution, and summary length. Our analysis showed that there were variations in the keyword frequency and summary length based on the language and news source.

In conclusion, our project demonstrated that combining translation and summarization techniques can lead to an efficient and effective system for processing news articles in different languages. The use of machine translation tools can be practical for large-scale translation tasks, while human translation can provide higher accuracy. The use of Spacy and BERT models for summarization proved to be effective in generating relevant summaries. Our findings can be useful in developing similar systems for news processing in the future.

Future Work

One area for future work could be improving the accuracy and performance of the translation and summarization techniques. While the results of the project were promising, there is still room for improvement, particularly in the accuracy of machine translation and the relevance of the summaries generated.

Another area for future work could be expanding the range of languages that the project can handle. While the project included several languages, there are still many other languages spoken around the world that could benefit from improved translation and summarization techniques.

In addition, the project could be expanded to include other types of texts beyond news articles, such as academic papers, legal documents, or social media posts. This would require adapting the techniques to the particular features and language used in each type of text.

Finally, another area for future work could be exploring the ethical implications of machine translation and summarization. As these technologies become more advanced and widely used, it is important to consider issues such as bias, accuracy, and the impact on human translators and interpreters.

References

Some of the references used are:

- Bhatia, A., Bansal, P., & Varma, V. (2020). A Survey on Cross-Lingual Summarization. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 19(4), 1-25. https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00520/114046/A-Survey-on-Cross-Lingual-Summarization
- Ganesan, K., Zhai, C., & Han, J. (2010). Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 340-348. https://www.aclweb.org/anthology/C10-2108.pdf
- Jain, S., & Chaturvedi, I. (2019). A survey of text summarization techniques. Journal of Information Processing Systems, 15(3), 539-557. https://www.jips-k.org/upload/pdf/1591JIPS_2019_v15n3_539.pdf
- Kulkarni, V. (2018). A Survey of Text Summarization Techniques. International Journal of Scientific & Engineering Research, 9(5), 1500-1505. https://www.ijser.org/researchpaper/A-Survey-of-Text-Summarization-Techniques.pdf

- Kulkarni, V., & Gujar, S. (2017). A survey of text summarization extractive techniques. International Journal of Computer Science and Information Technologies, 8(5), 2829-2832. https://www.researchgate.net/publication/322230842_A_Survey_of_Text_Summarization_Extractive_Techniques
- Maheshwari, R., & Anand, A. (2017). Automatic summarization techniques: A comprehensive survey. Artificial Intelligence Review, 47(4), 395-423. https://link.springer.com/article/10.1007/s10462-016-9491-8
- Saini, S., & Bansal, A. (2017). A review of extractive text summarization techniques. International Journal of Engineering and Computer Science, 6(6), 21611-21615. https://www.ijecs.in/index.php/ijecs/article/view/3937/3586
- Shukla, A., & Tripathi, A. (2019). A review of text summarization techniques. International Journal of Computer Science and Mobile Computing, 8(9), 48-53. https://www.ijcsmc.com/docs/papers/September2019/V8I9201910.pdf
- Tiwari, R., Singh, S., & Agarwal, S. (2016). A survey on text summarization techniques. International Journal of Computer Applications, 139(11), 19-24. https://www.ijcaonline.org/archives/volume139/number11/tiwari-2016-ijca-908735.pdf
- A Survey on Cross-Lingual Summarization Author: Ankush Gupta, Avik Ray, Amitava Das, Monojit Choudhury DOI: https://doi.org/10.1162/tacl_a_00520 Source: TACL Volume 8, 2020 - Issue
- Amit Kumar Verma, Prof. Pushpak Bhattacharyya, "A Literature Survey on Automatic Text Summarization", Indian Institute of Technology Bombay, March 2009, available at https://www.cfilt.iitb.ac.in/resources/surveys/amitkv_lit_survey_summarization.pdf

- Altmami, N. I., & Menai, M. E. B. (2020). Automatic summarization of scientific articles: A survey. Journal of King Saud University-Computer and Information Sciences, 32(4), 372-382. doi: https://doi.org/10.1016/j.jksuci.2019.12.002