Cross-Lingual Text Summarization

Manikrao Dhore
Student, Computer Engineering
Vishwakarma Institute of Technology
SPPU, Pune, India

Aaditya Ahire
Student, Computer Engineering
Vishwakarma Institute of Technology
SPPU, Pune, India

Aarya Mane
Student, Computer Engineering
Vishwakarma Institute of Technology
SPPU, Pune, India

Abubakar Bagban
Student, Computer Engineering
Vishwakarma Institute of Technology
SPPU, Pune, India

Samiksha Dongre
Faculty, Computer Engineering
Vishwakarma Institute of Technology
SPPU, Pune, India

***Abstract*—The process of creating a summary in one language (for example, English) for the provided document(s) in a different language (for example, Chinese) is known as cross-lingual summarizing. This task has gotten more and more attention from the computational linguistics community in the context of globalization. There is, however, still a need for a thorough examination of this assignment. As a result, we are presenting a comprehensive critical analysis of the datasets, methods, and difficulties in this area. We specifically arrange existing datasets and techniques in accordance with various building techniques and solution paradigms. Cross-lingual summarizing (CLS) tries to provide a summary of a document in a separate target language from the source language.**

Keywords—Cross-lingual

I. INTRODUCTION

Cross-Lingual Summarization (CLS) attempts to produce the matching summaries in a separate target language from texts in a source language. In an era of globalisation, CLS might make it easier for individuals to find important information in texts written in languages they don't understand. As a result, this task becomes increasingly crucial and is receiving significant study focus. Cross-Lingual Summarization (XLS), which creates a summary in the target language from the provided document(s) in a separate source language, attempts to make it easier for users to quickly understand the main points of documents written in a foreign language. This job might be viewed as a mix of machine translation (MT) and monolingual summarization (MS), both of which are tough natural language processing (NLP) problems that have been studied for decades .Due to its challenges and the lack of comparable corpus, XLS has received little attention in the statistical learning period despite its enormous practical value . The development of pre-trained encoder-decoder models has allowed neural summarizers and translators to perform .The two aforementioned results helped establish the XLS study area and slowly sparked interest in XLS. In example, during the past five years, academics have published more than 20 publications in an effort to solve the XLS task. However, a thorough examination of the advancements, difficulties, and potential of XLS is still lacking.

II. LITERATURE SURVEY

[1] They provide the first in-depth evaluation of current XLS research in this paper. We meticulously review the available XLS datasets and approaches, highlight their benefits, and contrast them with one another to provide deeper analysis. They also offer a number of viewpoints that might be taken into consideration for future research on XLS. The goal of this XLS survey is to enhance the state-of-the-art XLS technology while shedding light on the topic.

[2] In this article, they offer the MCLAS framework, a unique multi-task learning approach that enables cross lingual abstractive summarization with little parallel resources. Their methodology uses a single de cosequential generator to provide summaries in both one language and another. Experiments on two cross lingual summarization datasets show that our framework performs better than the baseline methods in both full-dataset and low-resource settings.

[3] In this technical research, they assess the cross lingual summarising zero-shot performance of popular bilingual and multilingual LLMs. They discover that the summarising and translating capabilities of Davinci-003, ChatGPT, and GPT-4 may be used to conduct zero-shot CLS and produce competitive results with the fine-tuned baseline. Additionally, the open-source LLMs that are now available often demonstrate their limited capacity to execute CLS in an end-to-end manner, demonstrating the difficulty of executing zero-shot CLS.

[4] The long-document cross-lingual summarization (long-document CLS) task is introduced in this study, and Perseus is suggested as the first long-document CLS dataset. On their dataset, they run numerous experiments and evaluate the benefits and drawbacks of various summarization techniques. They also offer an out-of-domain test set in the area of sports to assess the generalisation of long-document CLS models trained on our dataset. They personally examine the mLED summaries that were created, go further into the causes of these problems, and talk about potential remedies in order to better grasp the difficulties that Perseus poses. In the future, they want to concentrate on improving the Perseus multilingual version to accommodate the requirements of other languages and investigate a more effective approach for long-document CLS jobs.

[5] This paper introduces a new dataset for abstractive cross-lingual summarization of scientific papers in four languages: German, Italian, Chinese and Japanese. The paper also benchmarks different models based on a multilingual pre-trained model, and explores the benefits of intermediate-stage training using monolingual summarization and machine translation as intermediate tasks.

[6] This paper describes a study on cross-lingual summarization from English to Bahasa Indonesia, a language spoken in Indonesia. The study uses a direct cross lingual model without explicitly using a translator, thereby reducing one step as in existing method.

[7] The challenge of finding pertinent information when the document collection is written in a different language than the user query is known as cross-lingual information retrieval. A typical architecture for a CLIR system is shown in Figure 1 below. Due to the fact that the material is not in the user's native language, CLIR is frequently required.

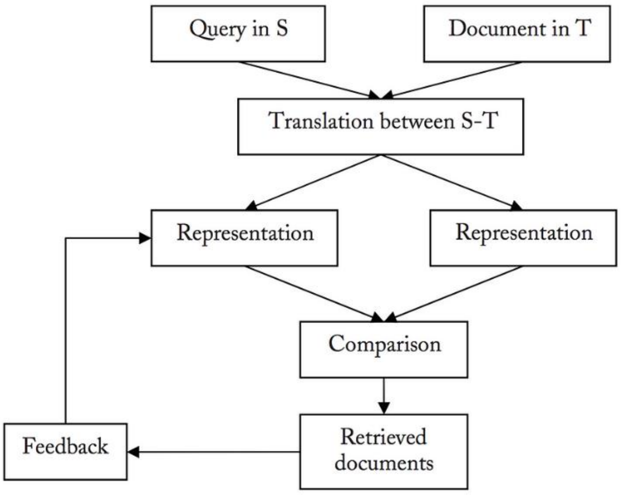


Fig 1. Typical Architecture of a CLIR System

[8] To obtain pertinent documents in CLIR, either the query, the document, or both must be translated into a common representation. The large resource requirements make it less attractive to translate all texts into the query language. In most cases, the query is converted into the language of the document collection. According to Jagalamudi and Kumaran (2008), machine readable bilingual dictionaries, parallel texts, and machine translation systems are typically utilised to translate the query. The majority of IR inquiries are brief in length and lack the essential syntactical components needed for machine translation. The majority of methods employ multilingual dictionaries to translate user queries. These multilingual dictionaries might not be complete. When a word's translation is unavailable, the word's transliteration is used instead.

[9] Recent developments in neural language models, such as BERT and XLNet, have produced some outstanding results in a variety of NLP tasks, such as phrase categorization, question answering, and document rating. In this study, they investigate how to model and learn the relevance between English queries and foreign-language publications in the task of cross-lingual information retrieval using the well-known bidirectional language model, BERT. By adjusting a pretrained multilingual BERT model with poor supervision using custom CLIR training data collected from parallel corpora, a deep relevance matching model based on BERT is introduced. The experimental results of retrieving Lithuanian materials using brief English queries demonstrate the efficiency of our model and its superiority over the competing baseline methods.

[10] In this study, they demonstrate a cross-lingual information retrieval (CLIR) system that, given a set of audio and text documents in a foreign language and a query in English, can return a scored list of pertinent documents and provide findings in an English summary form. A cutting-edge, multilingual, pretrained speech recognition model that is tailored to the target language first transcribing foreign audio material. For text documents, they employ numerous multilingual neural machine translation (MT) models to provide accurate translations, particularly for languages with limited or limited resources. The scores from a Neural Network Lexical Translation Model (NNLTM) and the likelihood of translation from GIZA translation tables are used to create a probabilistic CLIR model that assesses the processed texts and queries.

III. PROPOSED SYSTEM

1) The website serves as a graphical user interface (GUI) where users can input text and select their desired language for the summarized result.

2) Communication between the website and Python code is established through a Flask server, enabling data transfer and interaction.

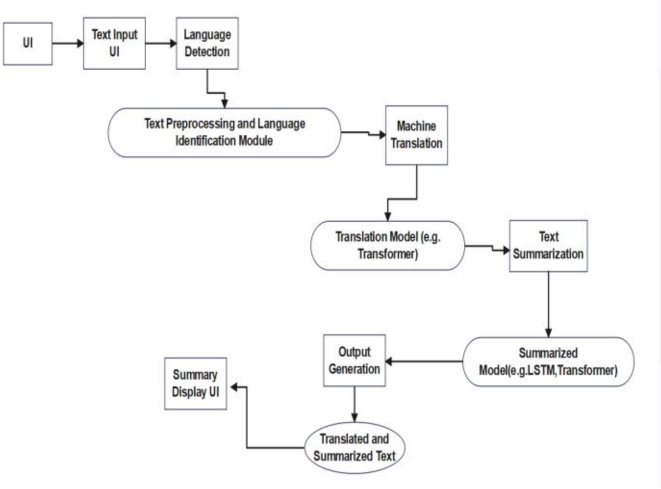
3) The Python program utilizes the Googletrans library to automatically detect the language of the input text.

4) The input text is then processed and transformed into English, preparing it for subsequent conversions.

5) The English equivalent of the user's input text is summarized using the 't5-small' model provided by the transformers library.

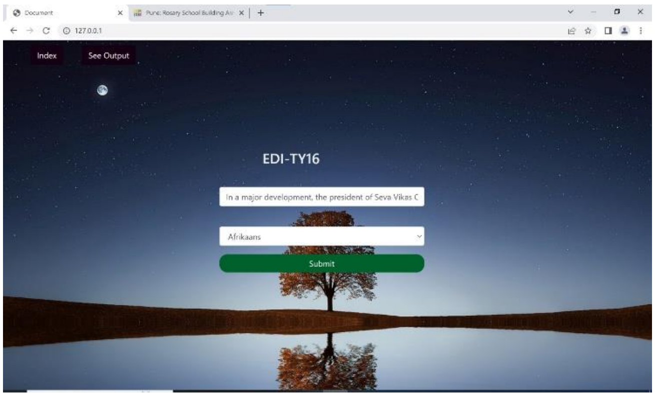
6) Finally, the summarized English text is translated into the user's selected language and displayed on the website, providing the desired outcome.

IV. SYSTEM ARECHITECTURE DIAGRAM


UI
Text Input UI
Language Detection
Text Preprocessing and Language Identification Module
Machine Translation
Translation Model (e.g. Transformer)
Text Summarization
Summarized Model(e.g.LSTM,Transformer)
Output Generation
Summary Display UI
Translated and Summarized Text

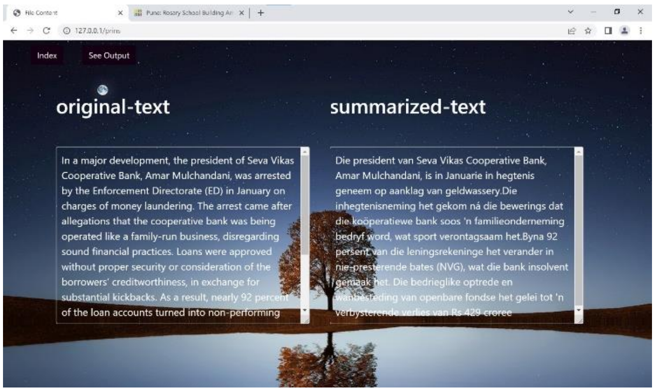
V. DESIGN AND IMPLEMENTATION

Given below are the snapshots of the website:



In a major development, the president of Seva Vikas C

Afrikaans



original-text

In a major development, the president of Seva Vikas Cooperative Bank, Amar Mulchandani, was arrested by the Enforcement Directorate (ED) in January on charges of money laundering. The arrest came after allegations that the cooperative bank was being operated like a family-run business, disregarding sound financial practices. Loans were approved without proper security or consideration of the borrowers' creditworthiness, in exchange for substantial kickbacks. As a result, nearly 92 percent of the loan accounts turned into non-performing

summarized-text

Die president van Seva Vikas Cooperative Bank, Amar Mulchandani, is in Januarie in hegtenis geneem op aanklag van geldwassery.Die inhegtenisneming het gekom ná die bewerings dat die koöperatiewe bank soos 'n familieonderneming bedryf word, wat sport verontagsaam het.Byna 92 persent van die leningsrekeninge het verander in nie-presterende bates (NVG), wat die bank insolvent gemaak het. Die bedrieglike optrede en wanbesteding van openbare fondse het gelei tot 'n verbysterende verlies van Rs 429 croree

VI. TECHNOLOGY STACK

1) Python- The general-purpose, interactive, object-oriented, and high-level programming language Python is particularly well-liked. Programming language Python uses dynamic typing and garbage collection.
2) Flask - Python-based Flask is a microweb framework. Due to the fact that it doesn't require any specific tools or libraries, it is categorized as a microframework. It lacks any components where pre-existing third-party libraries already provide common functionality, such as a database abstraction layer, form validation, or other components.
3) Google API (as translator)- With the use of cutting-edge Neural Machine Translation, the Translation API offers a straightforward, programmatic interface for translating any string into any supported language on-the-fly. In situations when the source language is unclear, it can also be used to identify a language.
4) HTML- The preferred markup language for Web pages is HTML. HTML may be used to create your own website.
5) CSS- The language we employ to style an HTML document is CSS.CSS outlines the presentation of HTML components.

VII. CONCLUSION

In conclusion, cross-lingual summarization is a challenging task that aims to generate concise summaries of documents in one language while targeting a different language. It involves several key components and techniques working together to achieve accurate and effective results. The system architecture for cross-lingual summarization typically includes a language detection module to identify the language of the input text, followed by text preprocessing and language identification steps. Machine translation is then used to translate the input text from the source language to the target language, leveraging powerful models such as Transformer architectures. After translation, a text summarization module generates a summary of the translated text, employing techniques like extractive or abstractive summarization using models such as LSTM or Transformers. The output generation stage provides the translated and summarized text, which is then presented to the user through a user interface or summary display. The overall process requires a combination of language processing, machine translation, and text summarization techniques, each tailored to handle the complexities of cross-lingual communication.

VIII. FUTURE SCOPE

Cross-lingual summarization is an area of research with significant potential for future advancements. Here are some potential areas of future scope for cross-lingual summarization:

1) Enhanced Multilingual Models: Current cross-lingual summarization systems often rely on machine translation models trained on parallel corpora. Future research can focus on developing more advanced multilingual models that can directly summarize documents in multiple languages without the need for translation. These models could leverage transfer learning techniques and shared representations across languages.
2) Fine-Grained Summarization: Cross-lingual summarization typically aims to generate a single summary for an entire document. Future work can explore fine-grained summarization techniques, where summaries are generated at the paragraph or sentence level. This approach can provide more

granular and context-specific summaries for different sections of the document.

3) Domain Adaptation: Cross-lingual summarization systems often face challenges when summarizing domain-specific content or documents from specific domains. Future research can focus on domain adaptation techniques to improve the performance of summarization models on specialized domains by leveraging domain-specific resources and pre-training.
4) Incorporating User Preferences: Personalization is an essential aspect of summarization systems. Future cross-lingual summarization models can be enhanced by integrating user preferences and feedback. This can enable the generation of summaries that align more closely with the specific needs and interests of individual users.
5) Multi-Document Summarization: Many real-world scenarios require summarizing multiple documents in different languages. Future research can explore techniques for multi-document cross-lingual summarization, where information is extracted and summarized from a collection of documents in various languages to provide a comprehensive and coherent summary.
6) Evaluating Cross-Lingual Summaries: Developing robust evaluation metrics and benchmarks for cross-lingual summarization is crucial for assessing and comparing the quality of different approaches. Future efforts can focus on creating standardized evaluation frameworks and datasets specifically designed for cross-lingual summarization, allowing for fair and reliable performance comparisons.
7) Low-Resource Languages: Cross-lingual summarization research has predominantly focused on widely spoken languages, while resource-scarce languages have received less attention. Future work can explore techniques for cross-lingual summarization in low-resource languages, where data and resources are limited, to facilitate information access and dissemination across diverse linguistic communities.

Overall, the future of cross-lingual summarization lies in advancing the state-of-the-art models, exploring new approaches for fine-grained and multi-document summarization, incorporating user preferences, addressing domain-specific challenges, and extending the scope to cover a broader range of languages and linguistic resources. These advancements will contribute to more effective and inclusive information summarization across language barriers.

REFERENCES

[1] A Survey on Cross-Lingual Summarization Jiaan Wang , Fandong Meng , Duo Zheng , Yunlong Liang Zhixu Li , Jianfeng Qu and Jie Zhou, 4Beijing University of Posts and Telecommunications,2022

[2] Cross-Lingual Abstractive Summarization with Limited Parallel Resources Yu Bai, Yang Gao, Heyan Huang, Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications, Beijing, China,2022

[3] Natural Language Processing based Cross Lingual Summarisation, Shree Akshaya AT, Shruthi Shanakaran, H M Thrupthi, Mamatha H R, IEEE(2022)

[4] Cross Lingual timeline Summarization, Luca Cagliero, Moreno La Quatra, Paola Garza, Elena Baralis, IEEE(2021) [5] Mary Jane C. Samonte, Renz A. Gazmin, John Derrik S. Soriano, Martela Nicolai O. Valencia “BridgeApp: An Assistive Mobile Communication Application for the Deaf and Mute”, 2019 International Conference On Communication and Techology Convergence,16-18 October 2019.

[5] X- SCITDLR :Cross-Lingual Extreme Summarisation of Scholarly Documents, Sotaro Takeshita; Tommaso Green; Niklas Friedrich; Kai Eckert; Simone Paolo Ponzetto, IEEE(2022)

[6] Cross Lingual Summarization:English- Bahasa Indonesia, Achmad F. Abka; Mahardhika Pratama; Wisnu Jatmiko, IEEE(2021)

[7] A Brief Introduction to Cross-Lingual Information Retrievalhttps://medium.com/lily-lab/a-brief-introduction-to-cross-lingual-information-retrieval-eba767fa9af6

[8] CLIR and its challenges https://www.cfilt.iitb.ac.in/resources/surveys/Swapnil-Cross-lingual-Information-Retrieval.pdf

[9] Cross-lingual Information Retrieval with BERT Zhuolin Jiang, Amro El-Jaroudi, William Hartmann, Damianos Karakos, Lingjun Zhao, Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)

[10] The 2019 BBN Cross-lingual Information Retrieval System ,Le Zhang, Damianos Karakos, William Hartmann, Manaj Srivastava, Lee Tarlin, David Akodes, Sanjay Krishna Gouda, Numra Bathool, Lingjun Zhao, Zhuolin Jiang, Richard Schwartz, John Makhoul, Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)