Developing Software That Automatically Translate Resource Materials between English to Indian Regional Languages

Mustafa Masuldar
Computer Science and Engineering
N B Navale Sinhgad College of Engineering, Solapur

Sakshi Deshmukh
Computer Science and Engineering
N B Navale Sinhgad College of Engineering, Solapur

Nandini Madre
Computer Science and Engineering
N B Navale Sinhgad College of Engineering, Solapur

Pooja Pawar
Computer Science and Engineering
N B Navale Sinhgad College of Engineering, Solapur

ABSTRACT

CIPAM, dedicated to advancing Intellectual Property Rights (IPR) awareness, commercialization, and enforcement, has created extensive educational materials. The software aims to bridge linguistic gaps by translating these materials from English to key Indian regional languages: Hindi, Marathi, Bengali, Gujarati, Tamil, and Telugu.

- The software's core features include its ability to translate various formats, such as Word documents, PDFs, and text within images. Notably, it focuses on maintaining the contextual integrity and
- Professionalism of the original content, ensuring that translations are not only accurate but also accessible to the general public.
- This abstract outlines a comprehensive approach involving Natural Language Processing (NLP) techniques, machine learning algorithms, and collaboration with experts to ensure the software's accuracy and user-friendliness.

Keywords: Research Paper, Technical Writing, Translation Software, Natural Language Processing (NLP), Engineering and Technology

I. INTRODUCTION

In today's globalized world, protecting Intellectual Property Rights (IPR) is crucial for driving innovation, economic growth, and sustainable development. Organizations like the Cell for IPR Promotion and Management (CIPAM) in India understand the importance of raising awareness about IPR and making it accessible to everyone. To tackle this challenge, CIPAM has developed advanced software that can translate IPR educational materials from English into major Indian languages like Hindi, Marathi, Bengali, Gujarati, Tamil, and Telugu.

This software is designed to handle various types of content, including text-based documents, PDFs, and even text within images. But what sets it apart is its ability to maintain the original meaning and professionalism of the content it translates. This ensures that the translated materials are not only accurate but also easy to understand for a wider audience, thus empowering more people with essential knowledge about IPR.

In this research paper, we'll take a closer look at CIPAM's multilingual translation software. We'll explore how it works, including the technical aspects like Natural Language Processing (NLP) techniques and machine learning algorithms. Additionally, we'll delve into real-world examples to understand how this software is making a difference in spreading awareness about IPR.

By examining the impact of CIPAM's innovative approach to translation, we hope to highlight the role of technology in promoting IPR awareness across diverse linguistic landscapes. Furthermore, we'll provide practical recommendations based on feedback

from users and stakeholders, aiming to improve similar initiatives and further advance the cause of IPR promotion in the digital age.

II. LITERATURE REVIEW

Adapted from the paper "***Jayanthi, N., Lakshmi, A., Raju, C. S. K., & Swathi, B. (2020). Dual Translation of International and Indian Regional Language using Recent Machine Translation. 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)"***

Title: *Dual Translation of International and Indian Regional Language using Recent Machine Translation*

Neural Machine Translation (NMT) is a modern and powerful approach that has resulted in major improvements compared to traditional techniques of machine translation for translating one language into another. In the world, India is a very multicultural and multilingual country. People of India from various regions use their own regional communication languages, making India stand at the second position in the world to have maximum languages. In India, English is provided as the second extra official language. But the usage of English in India is very less forming a communication gap. To minimize this gap by translating one language into another language is almost impossible for humans. It can be achieved by a machine translation. This paper focuses on translating Indic languages using the translation technique of Neural technology. A Sequence to Sequence model with encoder-decoder attention mechanism of neural machine translation is proposed for Telugu language conversion into English and vice versa.

Results of the paper ***"Nadeem Jadoon Khan, Waqas Anwar, Nadir Durrani, Machine Translation Approaches and Survey for Indian Languages"***

Title: *Machine Translation Approaches and Survey for Indian Languages*

In this study, we present an analysis regarding the performance of the state-of-art Phrase based Statistical Machine Translation (SMT) on multiple Indian languages. We report baseline systems on several language pairs. The motivation of this study is to promote the development of SMT and linguistic resources for these language pairs, as the current state-of-the-art is quite bleak due to sparse data resources. The success of an SMT system is contingent on the availability of a large parallel corpus. Such data is necessary to reliably estimate translation probabilities. We report the performance of baseline systems translating from Indian languages (Bengali, Guajarati, Hindi, Malayalam, Punjabi, Tamil, Telugu and Urdu) into English with average 10% accurate results for all the language pairs.

From this paper ***"Pulipaka, S. K., Kasaraneni, C. K., Sandeep Vemulapalli, V. N., & Mourya Kosaraju, S. S. (2019). Machine Translation of English Videos to Indian Regional Languages using Open Innovation. 2019 IEEE International Symposium on Technology and Society (ISTAS)."***

Title: *Machine Translation of English Videos to Indian Regional Languages using Open Innovation*

In spite of many languages being spoken in India, it is difficult for the people to understand foreign languages like English, Spanish, Italian, etc. The recognition and synthesis of speech are prominent emerging technologies in natural language processing and communication domains. This paper aims to leverage the open source applications of these technologies, machine translation, text-to-speech system (TTS), and speech-to text system (STT) to convert available online resources to Indian languages. This application takes an English language video as an input and separates the audio from video. It then divides the audio file into several smaller chunks based on the timestamps. These audio chunks are then individually converted into text using IBM Watson's speech-to-text (STT) module

From this paper ***"RAJANI S, Translation across Cultures: From the Regional to the Universal"***

Title: *Translation Across Cultures: From the Regional to the Universal*

India is a nation of many regional languages. The advent of the Europeans brought in a necessity to understand the regional languages through a

foreign language. The Europeans felt it necessary to learn and understand our languages so that their religion and religious teachings would reach the indigenous people better. In the process, many Christian missionaries who came to India learnt different dialects of the regional languages along with Sanskrit and made way for translation works. Reverend Ferdinand Kittle who came to India as a missionary and Ideologist, worked on similar lines and gave the monumental work; the kannada-English Dictionary. Even to this day this dictionary is considered as the authentic document of reference. This work helped many European readers to easily understand Kannada and hence connect themselves to the Kannada literary culture. In the long run numerous Kannada literary works of different genres were translated into English, giving it wider and broader readership. Here is a humble effort to peep into the translations of different genres of Kannada literature that have gained universal fame and applaud. Girish Karnad's Hayavadana, The Vachanas of Basavanna, Parva by S. L Bhyrappa, and few poems of K S Narasimha Swamy are a few works that are considered here.

From this paper ***"Sindhu, D. V., & Sagar, B. M. (2016). Study on machine translation approaches for Indian languages and their challenges. 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT). "***

Title: *Study on Machine Translation Approaches for Indian Languages and Their Challenges*

This survey mainly focuses on the developments of machine translation for the Indian languages. The survey throws a light on rule-based approach, empirical based approach and hybrid based approaches for machine translation. Every approach has its own advantages and disadvantages. Machine Translation (MT) is a process which translates from one language to another language. Due to rapid globalization there is an increased data over the web machine translation plays a very important role to reduce the language barrier between different regions. In a country like India with 22 official languages shows a high attention for the translation. This paper focuses on the different MT systems for Indian languages and also their challenges.

III. PROBLEM

Developing software to automatically translate resource materials between English to Indian regional languages.

IV. PROPOSED FEATURES

A. **Education and E-learning Platforms**: Implementing automatic translation software can significantly benefit educational institutions and e-learning platforms. It enables the translation of educational materials, textbooks, and online courses from English to various Indian regional languages, facilitating easier comprehension for students whose primary language may not be English. This application promotes inclusive learning and improves access to quality education.

B. **Government Services and Public Information:** Government agencies can utilize this technology to translate crucial information, public service announcements, forms, and documents into multiple Indian regional languages. This helps in disseminating important information to a broader audience, ensuring that all citizens can access and understand government services and policies effectively.

C. **Business and E-commerce:** E-commerce platforms and businesses aiming to expand their reach in India can benefit from automatic translation. Translating product descriptions, customer reviews, and marketing content into regional languages can attract and engage a more diverse customer base. This approach enhances user experience and increases sales by catering to non-English speaking populations.

D. **Healthcare Sector:** In the healthcare domain, translating medical resources, prescriptions, health guidelines, and patient information into regional languages can significantly improve healthcare accessibility. It assists healthcare providers in effectively communicating with

patients who might not be proficient in English, thus ensuring better healthcare outcomes and understanding.

E. **Content Creation and Media:** Content creators, publishers, and media houses can use automatic translation to reach a wider audience. Translating articles, news, blogs, and multimedia content into regional languages enhances audience engagement and expands the content's impact, allowing for broader dissemination and increased readership/viewership.

V. METHODOLOGY

WORKFLOW

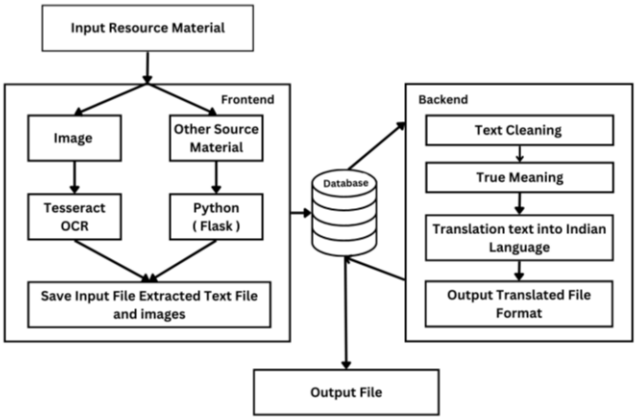

FIGURE 1: WORKFLOW DIAGRAM OF SOFTWARE

F. **Section Headings**

No more than 3 levels of headings should be used. All headings must be in 10pt font. Every word in a heading must be capitalized except for short minor words as listed in Section III-B.

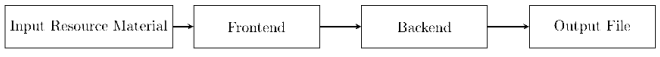

FIGURE 2: DFD LEVEL-0

1) DFD Level-0: The diagram shows the system at the highest level of abstraction, with a single input (image sent) and a single output (translated text).The system works by first extracting the text from the input image. This can be done using an optical character recognition (OCR) algorithm. Once the text has been extracted, it is cleaned to remove noise and errors. The cleaned text is then translated to the output language using a machine translation API. Finally, the translated text is saved to a file.

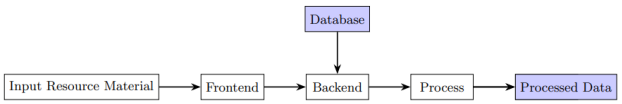

Figure 2: DFD Level-1

2) DFD Level-1: The image you sent is a flow chart for extracting text from a source material, such as a PDF, image, or video. The process can be divided into the following steps:

2.1) **Input**: The first step is to input the source material into the text extractor. This can be done by manually uploading the file or by using a URL to the file.

2.2) **Text Extraction:** The text extractor will then use various techniques to extract the text from the source material. This may involve OCR (optical character recognition) to extract text from images, or PDF parsing to extract text from PDF files.

2.3) **Text Processing:** Once the text has been extracted, it may need to be processed to clean it up and make it more machine-readable. This may involve removing non-textual elements, such as images and tables, or correcting spelling and grammar errors.

2.4) **Output:** The final step is to output the extracted text to a desired format, such as a text file, PDF file, or HTML file. This may involve additional processing, such as translation or summarization.

2.5) The flowchart can be adapted to extract text from other types of source material, such as images and videos, by using different text extraction techniques. The flowchart can also be adapted to include

additional steps, such as translation and summarization.

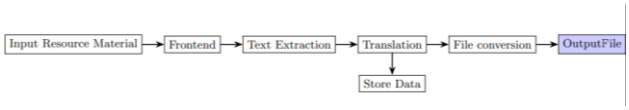

Figure 2: DFD Level-2

3) DFD Level-2: The DFD2 diagram you sent is a more detailed version of the Level 1 DFD for a Text Extraction System. It breaks down the Text Extractor, Text Cleaner, and Translator subsystems into their components and data flows.

- **Text Extractor:** The Text Extractor subsystem is responsible for extracting text from a source material. It has the following components:
- **OCR (Optical Character Recognition) component:** Extracts text from images. PDF Parsing component: Extracts text from PDF files.
- **Other Text Extractors**: Extracts text from other types of source material, such as websites, emails, and social media posts.

The data flows between the Text Extractor components are as follows:

1. **Source Material to OCR component:** This data flow contains the source material that the user wants to extract text from.
2. **OCR Output to PDF Parsing component (optional**): This data flow contains the extracted text from the OCR component. It is optional because the PDF Parsing component can also receive text from other components.
3. **Other Text Extractors output to PDF Parsing component (optional**): This data flow contains the extracted text from the other text extraction components. It is optional because the PDF Parsing component can also receive text from other components.
4. **PDF Parsing output to Text Extractor output**: This data flow contains the extracted text from the PDF Parsing component.
5. **Text Cleaner:** The Text Cleaner subsystem is responsible for cleaning up the extracted text to make it more machine-readable.

Sequence Diagram of software

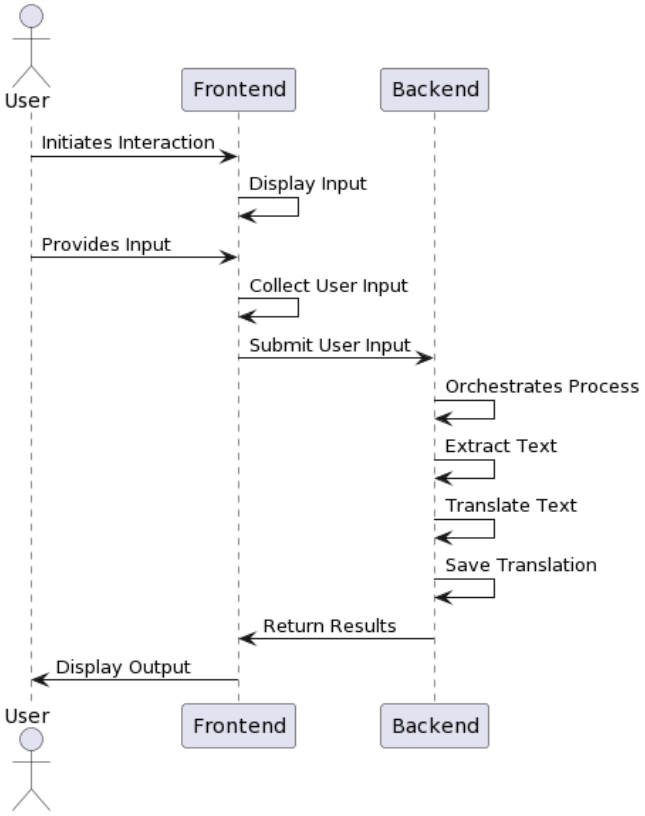

Figure 2: Sequence Diagram of software

The sequence diagram you provided shows the steps involved in extracting text from an image using a distributed system. The system consists of the following components:

1) **User**: The user sends an input image to the system.
2) **Frontend:** The frontend receives the input image from the user and forwards it to the backend.
3) **Backend:** The backend coordinates the text extraction process. It requests the OCR service

to extract the text from the image, and then pre-processes and cleans the extracted text. Finally, it saves the cleaned text to the database and requests the Output File Generator component to generate an output file.

VI. SYSTEM TESTING

The system testing for this project involved rigorous evaluation of the software's functionality and performance on a web application. The testing process was designed to verify the application's compatibility with the specified device and ensure seamless operation across various features. Notably, particular emphasis was placed on assessing the software's ability to translate educational materials from English to key Indian regional languages, including Hindi, Marathi, Bengali, Gujarati, Tamil, and Telugu. Additionally, the testing phase scrutinized the software's capability to handle diverse file formats such as Word documents, PDFs, and text within images, while maintaining accuracy and accessibility for the general public. It's essential to highlight that the data retrieval speed from the internet was considered, contingent upon the network bandwidth provided by the user's service provider. Nonetheless, the prevalence of 4G/5G mobile networks in India was leveraged to ensure faster data access and enhance user experience.

VII. REFERENCES

[1] N. Jayanthi, A. Lakshmi, C. S. K. Raju, and B. Swathi, "Dual translation of international and Indian regional language using recent machine translation," in 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), December 2020, pp. 682-686.

[2] A. Godase and S. Govilkar, "Machine translation development for Indian languages and its approaches," International Journal on Natural Language Computing, vol. 4, no. 2, pp. 55-74, 2015.

[3] N. J. Khan, W. Anwar, and N. Durrani, "Machine translation approaches and survey for Indian languages," arXiv preprint arXiv:1701.04290, 2017.

[4] A. Kunchukuttan, A. Mishra, R. Chatterjee, R. Shah, and P. Bhattacharyya, "Sata-anuvadak: Tackling multiway translation of Indian languages," pan, vol. 841(54,570), pp. 4-135, 2014.

[5] S. K. Pulipaka, C. K. Kasaraneni, V. N. S. Vemulapalli, and S. S. M. Kosaraju, "Machine translation of English videos to Indian regional languages using open innovation," in 2019 IEEE International Symposium on Technology and Society (ISTAS), November 2019, pp. 1-7.

[6] S. Rajani, "Translation across Cultures: From the Regional to the Universal," Think India Journal, vol. 22, no. 6, pp. 189-194, 2019.

[7] S. K. Dwivedi and P. P. Sukhadeve, "Machine translation system in Indian perspectives," Journal of Computer Science, vol. 6, no. 10, pp. 1111, 2010.

[8] J. Ray, "A review of terminological work being done in Indian Languages," in Proceedings of Translating and the Computer: Term banks for tomorrow’s world, 1982.

[9] D. V. Sindhu and B. M. Sagar, "Study on machine translation approaches for Indian languages and their challenges," in 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), December 2016, pp. 262-267.

[10] M. Singh and P. Bhatia, "Automated conversion of English and Hindi text to Braille representation," International Journal of Computer Applications, vol. 4, no. 6, pp. 25-29, 2010.