International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 04 | Apr 2023 www.irjet.net p-ISSN: 2395-0072

Translation Ally: Document and Audio Translator

Zalak Gandhi[1], Saloni Joshi[2], Mansi Kargutkar[3], Khushi Pal[4] ,Renuka Nagpure[5]

[1,2,3,4] *Undergraduate, IT Engineering, Atharva College of Engineering, Maharashtra, India*
[5]*Professor, IT Department, Atharva College of Engineering, Maharashtra, India*

Abstract - *For years, language has been a barrier for many companies, and people especially for companies and employees, many companies cannot extend their businesses, and many employees are not able to work in specific countries and specific companies, just because of different languages. Obstacles or issues that prevent information from flowing between a sender and a receiver cause the communication process to fail and are referred to as barriers to effective communication. If someone's words don't make sense to us, every conversation, email, report, and memo will be unproductive. Simple daily tasks might be made difficult by language limitations. As more businesses move overseas, linguistic hurdles may become a worldwide problem. This creates a hindrance in conducting business smoothly. Our project will remove language barriers in business communication worldwide, it will be advantageous for multinational companies and businesses all over the world. Our project will help to extend their business overseas without any language hindrance. Our project will be a platform where one can convert all types of text documents, audio, and video transcripts to any other language. We make use of python libraries and django for translating documents and audio files. This web application will prevent miscommunications, misunderstandings, and conflicts. It will convey thoughts, ideas, and instructions more effectively.*

Key Words: **Translation, document, audio, python, django.**

1. INTRODUCTION

Human beings experience a language barrier when they are unable to communicate using a particular language. There are several causes of language barriers. When people speaking in different languages interact with each other, they do not understand each other so there is no point of communication. According to some statistics, if more than 10,000 people speak 121 different languages throughout the nation, it is not necessary for someone else to be able to understand that specific language. English is not the first language for most of the people in the world. Therefore, it is essential to have a translator.

The conversation is meaningless if the speaker and the recipient do not use the same words and language. Communication can become ineffective and messages may not get across if certain words are not used that the other person can comprehend.

Language barriers can be challenging to surmount, whether you're traveling and attempting to understand a restaurant menu or working in a multilingual company. And work is no different. In reality, being unable to accurately follow a conversation in a business situation only makes it more scrutinized and potentially embarrassing. The revolution in remote work has enabled companies to look beyond their physical borders and penetrate untapped markets. This has caused a sudden revelation to sweep the business world.

Translation Ally will enable you to translate any kind of document and audio in any language to any other language and in turn remove the language barriers that are caused in businesses. Suppose an employee is transferred from one state or country to another in this case they can make use of this website and translate the documents and audio which are in a language that is unknown to them to a language which they can understand.

Considering this problem of language barrier in business communication, our main aim of the project is to create a web application that can overcome language barrier problems around the globe. To help bridge the language barrier, document translators will effectively translate business reports, excel sheets, letters, etc from one language to another language without losing the formatting of the document. The employees who have to visit other countries for work purposes and don't know the language of that country can use our website to translate all the work related documents and instructions provided to them. They don't have to specifically learn a foreign language to communicate with their colleagues. Our project will help them to adjust to the new environment and make them feel at ease.

The solution translates different formats of documents such as .txt,.docx,.xlsx,.csv and audio files such as mp3 are translated as per users' needs. First the user will have to

upload the file or audio and select the source language and desired language.

1.1 Software Overview

Our project will help in translating different formats of documents used in businesses. Users will get an interface/home page where the user can upload a document and also gets the option to select languages. Users will have to select the source language in which the original document is written. Then the user will have to select the destination language viz in which they want the output of the uploaded document. Once a user uploads their file and selects the languages, the web application generates a translated file for the user. Files can be of various forms such as docx, excel, csv, mp3 etc. For docx files the translated file will preserve all the contents and styling of the original file like tables, images and fonts. After a few seconds or minutes, the user will get a pop up message if the file is ready for download, and after that the user can click on the download button given and the file will be ready to download.

2. LITERATURE SURVEY

1) The web application, Document Segmentation and Language Translation Using Tesseract-OCR by S. Thakare, A. Kamble, V. Thengne and U. R. Kamble. This paper explains an application that accepts image documents as an input, a user defines an image file containing text in any language available in the Python-tesseract library and does its translation in any supported languages using Google Translator.

2) An overview of The impact of language barriers on businesses by Helene Tenzer, Markus Pudelko & Anne-Wil Harzing. This paper discusses that as corporations extend their businesses over the seas there is hindrance of language. They need to overcome this barrier in order to make their business successful.

3) Language translation of web-based content by Bart Kalher, Brain Bacher, K. C. Jones. This paper summarizes a project that can translate websites and help people surf the web without any boundaries. It provides adequate conversion of foreign languages to one's native tongue; however, dialects, slang, and character conversion errors result in partially successful translations.

4) An efficient English to Hindi machine translation system using hybrid mechanism by J. Nair, K. A. Krishnan and R. Deetha. This paper discusses that as the majority of Indians, especially those living in distant villages, cannot read, write, or understand English, an effective language translator must be used.

2.1 Comparative Analysis

In this section a comparison between the existing system and proposed system is conducted.

Table-1: Comparative Analysis

Current System	Proposed System
It supports upto 67 languages.	It supports upto 107 languages.
System is not able to preserve formatting.	System preserves formatting.
Audio translation is not included in the system	Audio translation is implemented

3. SURVEY OF TYPES OF DOCUMENTS USED

According to a survey conducted us, there are three common document formats used in business:

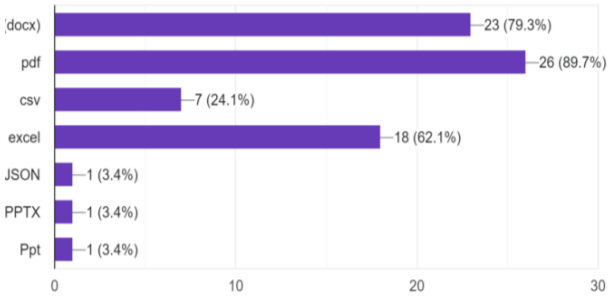

Fig -1: Most used format used in businesses

The survey was conducted in August 2022. The survey was conducted by sharing google forms with the people currently in co-operates and working in different industries.

This survey form contained a few questions like which documents their company uses for different documentation purposes and for legal documents. As per the responses, it was found that the top three documents are .pdf, .docx and .xlsx format.

PDF (Portable Document Format): This format is widely used in business for documents such as contracts, proposals, and reports. PDFs are widely supported and

maintain formatting across different devices and operating systems.

Microsoft Word: Word is the standard word processing program used by many businesses. It allows for easy collaboration and formatting, and can be used for a variety of documents such as letters, memos, and reports.

Excel (or CSV): Excel is used for spreadsheets and data analysis. It can be used for financial analysis, budgeting, and inventory management. CSV (Comma-Separated Values) is a simpler format used for exporting and importing data between different programs.

We have tried to include the translations for these documents in our project. This will in turn help the users in translating important documents like contracts, proposals, and reports.

4. TECHNOLOGIES USED

This project was developed by using the combination of various technologies, languages and algorithms which made it to where the optimum functionality was achieved.

4.1 Python-Django

Python is an open source, extensible, scalable, and heavy processing capable programming language that we used to develop the backend of our application. As it supports various libraries and features, this was the best choice we can make. With reference to that, the very popular framework of Python i.e. Django was used. This is a framework that helps the process of developing web related applications easy and hard-code free.

4.2 HTML, CSS and JavaScript

HTML is a hyper-text markup language. It is for designing web pages. On the World Wide Web, it is utilized for material presentation and structuring. It combines scripting languages like JavaScript and tools like Cascading Style Sheets (CSS).

5. PYTHON LIBRARIES USED

Python supports a variety of libraries to ease programmers work and provide a wide range of functionality.

5.1 BeautifulSoup

The Beautiful Soup Python library is used to analyze HTML and XML texts. For parsed sites, it generates a parse tree that can be used to extract HTML data for web scraping. In this project, it is used for reading an html document file which was a .docx file converted to HTML, for preserving the styling and format while translation.

5.2 NLTK

The Natural Language Toolkit, or more simply NLTK, is a collection of Python-coded tools and applications for natural language processing of English.

5.3 Mammoth

It is one of the libraries used to convert .docx formatted files into HTML files. It is used while translating .docx, to convert it to HTML for preserving fonts, styles, colors, tables and formatting in a document and giving it precisely back to the user.

5.4 Pandas

It is one of the powerful packages of python for handling data. It is a simple, fast, expressive library. It is used in this project for reading and writing of .csv and .xlsx files.

6. SYSTEM FLOW

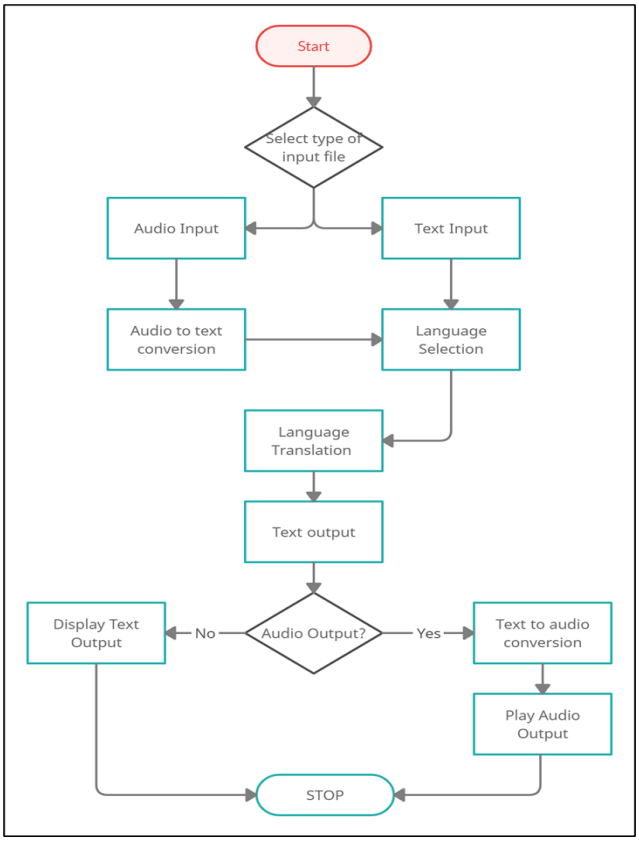

Fig -2: Flowchart

This Web application will help in translating different formats of documents and audio such as docx, excel, csv, mp3 etc used in businesses and in any other field.

1. The Users have to select a document which they want to translate

2. The users will have to select the source language in which the original document is written. Even if the user does not select the source language or selects incorrect language then it will automatically detect the source language. Then the user will have to select the destination language in which they want the uploaded document to be translated.

3. The uploaded file can be an audio file or a document. If the uploaded file is an audio file then text will be extracted from the audio file and will be stored in a text file.

4. The translation of the text will take place with the help of googletrans library and output will be generated.

5. According to the needs of the user, the output will be provided in audio or text format. For docx and pdf files the translated file will preserve all the contents and styling of the original file like bold, underline, tables, images and fonts.

6. METHODOLOGY

When a user uploads a document the document gets stored in the backend. Users can upload different types of files like docx, excel, csv, text. Depending on the type of file there are different methods for processing and extracting text from the uploaded file.

6.1. Text File

For text files the file is opened and read using normal file handling functions provided by python. The sentences are separated using the nltk library and are stored in a list. The list is iterated and each sentence is translated one by one. The translated sentences are overwritten in the same file.

6.2. Docx File

The docx file is converted to an html file. This conversion is done to preserve the formatting of the docx file, so that while translating the file the images, tables and various other stylings are not lost. The images in the docx file are encoded to base64 format so that they can be recreated when the file is converted back to the original format. The conversion of docx file to html is done by a library called mammoth. The contents of the docx file are wrapped into various tags and the text in the <p>, <li> , <td>,etc tags is extracted and is translated one tag at a time. The parsing of the html file is done using Beautiful soup library. It will create a parse tree for all parse pages that can be used to extract data from HTML, which is useful for web scraping. The original text is replaced with the translated text and in this way all the text in the original file is replaced with its corresponding translation. From the translated html file a docx file is created containing the same format as that of the uploaded file. The base64 images are decoded and the original image is obtained.

6.3. Csv File

Csv files are comma separated files which are generally used to store a large amount of data. The csv files are read as normal text files, since python takes the least amount of time to process text files. So to decrease the response time and increase the efficiency of the website the csv files are read using python file handling functions. Each row is taken from the csv file and given as an input to the translator function. The translated text is overwritten in the same file.

6.4. Excel file

Excel sheets are most commonly used in corporations for storing business data. Excel files are generally larger in size compared to csv files and hence take a lot of processing time. So to decrease this processing time excel files are converted to csv files which are we can say a compressed version of the excel files. The conversion of the excel file to csv is done using pandas library. The csv file is then translated one row at a time and original rows are overwritten by the translated rows. After the csv file is translated it is converted back to excel file.

7. RESULT

This is the document in english language which we uploaded in the project for translation.

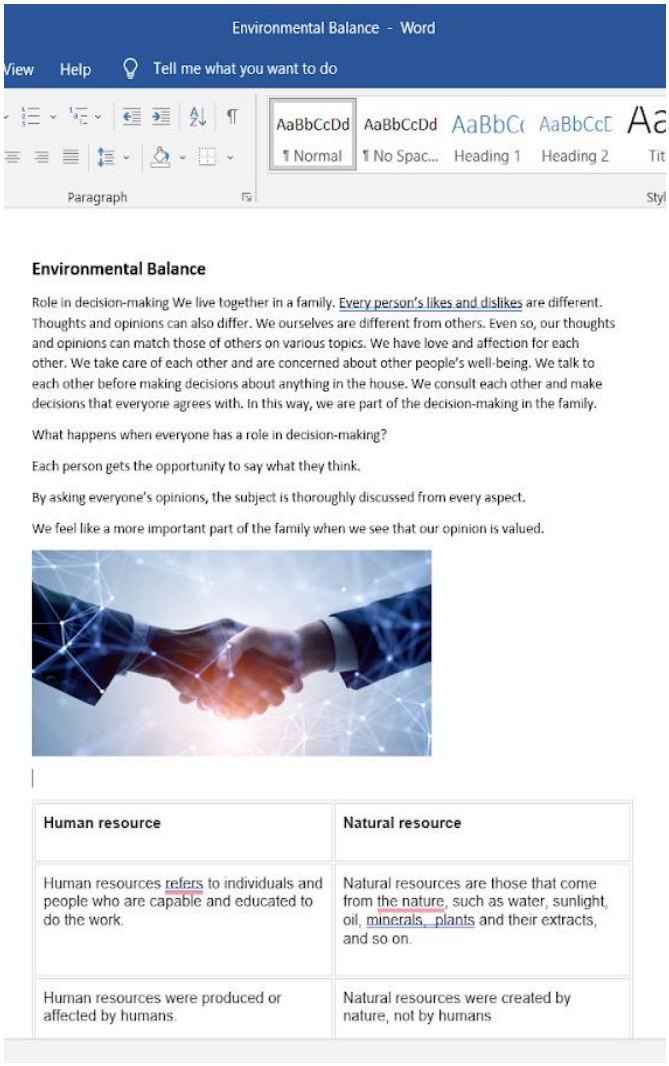

Fig -3: Uploaded Document(Before Translation)

Below is the translated document which got downloaded and got translated in Hindi with all the styling, images and tables preserved with formatting.

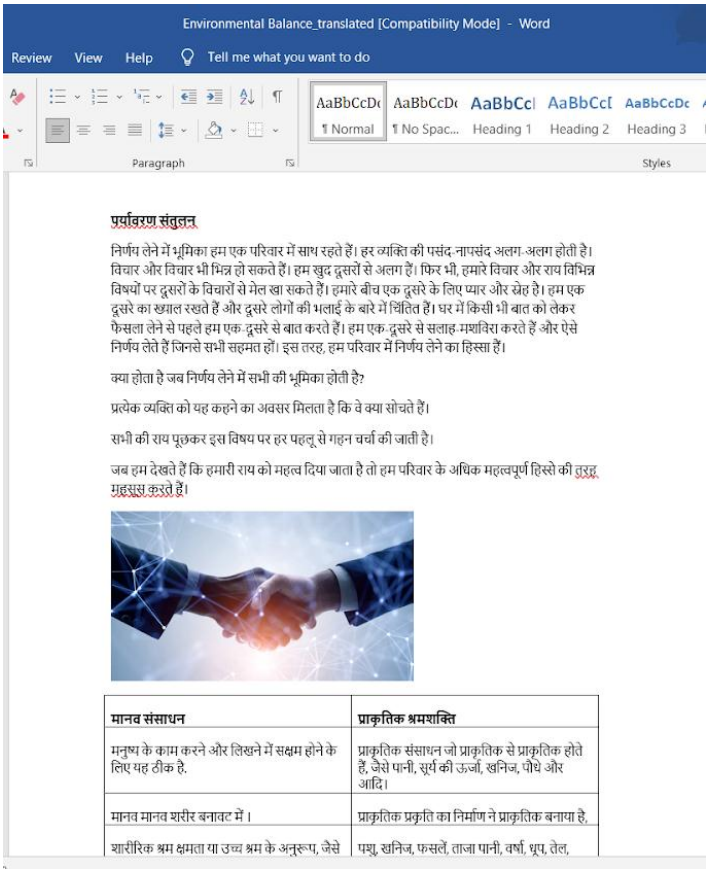

Fig -4: Downloaded Document(After Translation)

8. CONCLUSIONS

Language can result in the biggest barrier in the business world and can also create miscommunication and misinterpretation among the workers. Each and every business wants to grow and expand across the globe and as different languages are used in different states and countries so to understand a particular file or document the user does not need to learn that particular language instead can use this project to translate the file.

This is very simple and easy to use for anyone, the user will get the translated file in a few clicks and inputs. The user can get the translated files in a few seconds and for free. File can be of different formats such as .txt, .docx, .csv, .xlsx, etc and also the translated file will be of the same format. In simple words document translation is the process of converting the text from one language to another. Depending on the industry in which the user operates, any number of documents and content can require translation. Other than business documents in different fields such as healthcare, government, law and in many others it can be used to translate the file, the benefit of this document translation is very vast.

ACKNOWLEDGEMENT

We would like to express our profound gratitude to Prof Deepali Maste, HOD of Information Technology Department, and Dr Shrikant Kallulkar Principal of Atharva college of engineering for their contributions to the completion of our project titled Document Translator.

We would like to express our special thanks to our project guide and Major project co-ordinator Prof Renuka Nagpure for her time and the efforts she provided throughout the year. Your useful advice and suggestions were really helpful to us during the project's completion. In this aspect, we are really grateful to you.

REFERENCES

[1] S. Thakare, A. Kamble, V. Thengne and U. R. Kamble, "Document Segmentation and Language Translation Using Tesseract-OCR," 2018 IEEE 13th International Conference on Industrial and Information Systems (ICIIS), 2018, pp. 148-151, doi: 10.1109/ICIINFS.2018.8721372.

[2] B. Kahler, B. Bacher and K. C. Jones, "Language translation of web-based content," 2012 IEEE National Aerospace and Electronics Conference (NAECON), 2012, pp. 40-45, doi: 10.1109/NAECON.2012.6531026.

[3] J. Nair, K. A. Krishnan and R. Deetha, "An efficient English to Hindi machine translation system using hybrid mechanism," 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2016, pp. 2109-2113, doi: 10.1109/ICACCI.2016.7732363.

[4] Pudelko, Markus, and Helene Tenzer. "Boundaryless careers or career boundaries? The impact of language barriers on academic careers in international business schools." *Academy of Management Learning & Education* 18.2 (2019): 213-240.

[5] U. Kheradia and A. Kondwilkar, "Speech To Speech Language Translator", International Journal of Scientific and Research Publications, Volume 2, Issue 12, December 2012 1 ISSN 2250-3153.

[6] A. Waibel, A. Badran, A. W Black, R. Frederking, D. Gates, A. Lavie, K. Lenzo, L. Tomokiyo, J. Reichert, T. Schultz, D. Wallace, M. Woszczyna and J. Zhang, "Speechalator: two-way speech-to-speech translation on a consumer PDA", EUROSPEECH 2003 – GENEVA.

[7] B. Turovsky, "Found in translation: More accurate, fluent sentences in Google Translate", Published Nov 15, 2016.

[8] R. Sennrich, B. Haddow and A. Birch, "Neural Machine Translation of Rare Words with Subword Units" , Submitted on 31 Aug 2015 (v1), last revised 10 Jun 2016 (this version, v5)) , The research presented in this publication was conducted in cooperation with Samsung Electronics Polska sp. z o.o. -Samsung R&D Institute Poland.

[9] Prof.N.R.Ingale, Ashish Suman, Aniruddha Patil, Suhasini Raina, "Text Fetching App by Image Processing" Published in International Research Journal of Innovations in Engineering and Technology , Volume 4, Issue 5, pp 51-54, May 2020.

[10] G. Lample, L. Denoyer and M. Ranzato, "UNSUPERVISED MACHINE TRANSLATION USING MONOLINGUAL CORPORA ONLY", Under review as a conference paper at ICLR 2018.

BIOGRAPHIES

Miss Zalak Gandhi, resident of Mumbai, Maharashtra, India, Student of IT Engineering from Mumbai University,

Miss Saloni Joshi, resident of Mumbai, Maharashtra, India, Student of IT Engineering from Mumbai University,

Miss Mansi Kargutkar, resident of Mumbai, Maharashtra, India, Student of IT Engineering from Mumbai University,

Miss Khushi Pal, resident of Mumbai, Maharashtra, India, Student of IT Engineering from Mumbai University,