International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Conference Proceedings: Volume: 05 ACME – 2026 DOI: 10.55041/ISJEM.ACME086
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Automated Multilingual Translation of Documents and Question Papers using Generative AI Techniques

Ms.S SHAHEEN [1]
Assistant Professor, Department of AI&DS
Annamacharya Institute of Technology and Sciences, Tirupati – 517520, A.P.
shaheenshaalu@gmail.com

S YUGANDHAR KUMAR [2]
UG Scholar, Department of AI&DS
Annamacharya Institute of Technology and Sciences, Tirupati – 517520, A.P.
yugandharsirigiraju8@gmail.com

P RAHUL REDDY [3]
UG Scholar, Department of AI&DS
Annamacharya Institute of Technology and Sciences, Tirupati – 517520, A.P.
rahulreddyreddy789@gmail.com

R SAI KISHAN VARMA [4]
UG Scholar, Department of AI&DS
Annamacharya Institute of Technology and Sciences, Tirupati – 517520, A.P.
saikishanvarma.5@gmail.com

P BALA VAMSI KRISHNA [5]
UG Scholar, Department of AI&DS
Annamacharya Institute of Technology and Sciences, Tirupati – 517520, A.P.
krishnabalavamsi8@gmail.com

***Abstract*—** The increased demand for multilingual education materials has posed a major challenge for institutions that require translating examination question papers in various regional languages. Traditional translation methods involve a significant amount of time and resources and may also introduce inconsistencies in the translation process, affecting the clarity of the examination process. This paper proposes an automated multilingual document translation system using Generative AI and Neural Machine Translation (NMT) approaches for the efficient translation of examination question papers in various languages. In the proposed system, the translation document can be in either PDF or DOCX formats. Text content is extracted using the pdfplumber and python-docx libraries for PDF,Text and Word documents, respectively. A Google Neural Machine Translation model is employed using the Deep Translator API for translation. This system is implemented using Python and the Django web development framework. This provides a simple and effective interface for uploading the document and generating the translation output. Observations made on the proposed system indicate the efficient translation of documents in various contexts, particularly in the field of education. This makes the proposed approach more suitable and effective in the field of education.

Keywords: Document Translation, Generative AI, Neural Machine Translation, Natural Language Processing, Deep Translator API, Document Processing, Text Extraction, pdfplumber, python-docx, Django Framework, Educational Technology, Automated Question Paper Translation.

I.INTRODUCTION

The rising need for multilingual educational resources has posed great challenges for educational institutions that need to prepare examination question papers in various regional languages. In multilingual countries, it is essential to provide students with examination resources in various languages to ensure fairness in the examination process. Traditionally, the process of translating the documents into various languages is a time-consuming process, which not only incurs significant costs but also introduces translation errors that alter the original meaning of the educational questions. Such translation errors affect the clarity of the examinations, which in turn affects the evaluation process [1]-[3].

Most of these organizations rely on human translators or general online translation tools to translate their documents

into different languages. Though human translation is reliable, it is a tedious and expensive task, especially when one has to translate a large number of documents within a given period. Similarly, general online translation tools require a lot of copying and pasting of texts, and they do not maintain document structures, formats, numbering, and academic vocabulary. These challenges associated with current translation techniques make them inefficient in dealing with large academic documents such as exam papers and study materials [4], [5].

Recent developments in Artificial Intelligence (AI) and Natural Language Processing (NLP) technologies have made it possible to develop intelligent automated translation systems that can translate text from one language to another and maintain context. For instance, a class of machine translation technologies known as Neural Machine Translation (NMT) uses deep learning techniques that analyze a sentence as a whole, thus allowing for more effective semantic relationships and linguistic patterns. As a result, it is more suitable for multilingual document processing compared to traditional machine translation techniques, since it is more fluent and contextually accurate in translating text [6]-[8].

Motivated by the above innovations, this paper proposes the Automated Multilingual Document Translation System using Generative AI and Neural Machine Translation. The proposed system will translate the academic documents and the question papers using the Automated System. The proposed system will accept the document in PDF format, extract the text using the Document Processing Libraries, perform the pre-processing operations, translate the text using the Google Neural Machine Translation Model via the Deep Translator API, reconstruct the translated text in the required format, and allow the user to download the translated document in the required format. The proposed framework combines the Document Parsing, Natural Language Processing, and AI-based translation techniques to provide the efficient solution for the preparation of the multilingual academic documents [9], [10].

II. RELATED WORK

The development of automated language translation systems has been able to attract a lot of attention in recent times due to the growing need for multilingual communication using digital mediums. Initially, automated language translation systems were developed using machine translation techniques that relied on pre-defined grammar and dictionary rules. These translation systems were highly dependent on grammar rules. These translation systems were not effective in translating complex sentences. Therefore, the results obtained using this translation system were quite rigid and lacked flexibility.

The problem of translation in a multilingual document has been addressed in various research works done in natural language processing and machine translation. In past research works done to address the translation problem in a document, machine translation and statistical machine translation approaches have been used. In both of these approaches, various rules are used to translate a text from one language to another. Though this approach has laid a foundation for translation, context is not preserved.

With the advent of computational linguistics, it has been possible for researchers to examine the role of applying statistical techniques for improving the quality of translation. Statistical Machine Translation models involve the study of a large amount of data in two languages in order to compute the probability of translating a word or a phrase. Even though it has been seen that the accuracy of translation is high in comparison to rule-based translation systems, it has been a challenge for statistical models to maintain consistency in translation while translating sentences.

However, with the recent advancements in the field of Artificial Intelligence and Deep Learning, Neural Machine Translation models have been developed. Neural Machine Translation models are made of a neural network that has the ability to learn the semantics of words in a language and thus has the ability to translate a sentence while keeping the context of all the sentences. It has been observed that Neural Machine Translation models have improved fluency and accuracy in translation. However, there are a number of issues related to the implementation of the translation models in structured documents such as academic and examination papers.

Parallel to this, the management of different document formats has also become an essential part in the development of translation systems. Question papers in various educational institutions are mostly in formats like PDF and DOCX. It is necessary to extract the documents

effectively before the translation process is applied. pdfplumber and python-docx are some of the most widely used tools for effectively extracting the documents before the translation process is applied.

Though the state-of-the-art translation tools and the document processing tools are already in place, there is a need for the development of a simple and unified system that can effectively integrate the tools in a unified manner. However, the existing tools are not effectively addressing the issue of translating structured educational documents in a way that can be utilized by the institutions. This system aims at effectively addressing the issue by integrating the tools in a unified manner.

III. METHODOLOGY

This section presents the methodology used in the development of the proposed multilingual document translation system. The proposed system is designed to simplify the translation process of educational documents, especially question papers for examinations, into multiple languages. The system combines the process of document handling, text pre-processing, Neural Machine Translation, and document generation into one process.

A. Document Input and Data Preparation

The proposed system is designed to accept input documents in the most commonly used formats, namely PDF and DOCX. These formats are the most widely used in the preparation of educational materials. The educational materials usually contain structured information such as questions, guidance, etc. The proposed system will accept the input document in the required format through the web interface. Then, it will identify the format of the input document and extract the text accordingly. In the case of a PDF document, the pdfplumber library will be used to extract the text. Similarly, in the case of a DOCX document, the python-docx library will be used to extract the text.

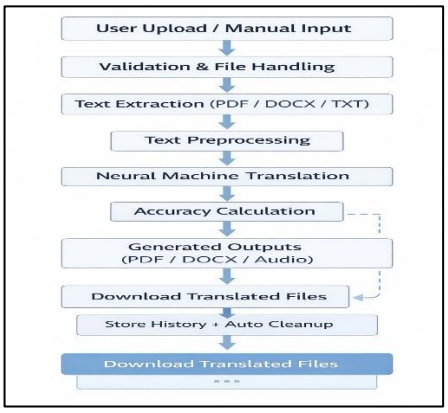

Fig. 1. System architecture of the automated multilingual document translation system.

This figure shows a general overview of how the system works, including how a document is uploaded, text is extracted, preprocessed, translated, and then output.

B. Text Preprocessing

In this stage, text is preprocessed for better translation output. It is important to note that text is preprocessed for better translation output. It is important to note that text is divided into smaller units, such as sentences, to ensure that it is translated more effectively. Some of the text preprocessing techniques include the removal of spaces, text formatting, and structuring of text.

C. Neural Machine Translation

The translation process occurs through the use of Neural Machine Translation technology. In this regard, the system uses the Deep Translator API, which in turn uses the Google Neural Machine Translation model internally. The translation model receives the text segment as well as the chosen translation language. The translation process occurs by the translation model processing the text segment and producing the translated output.

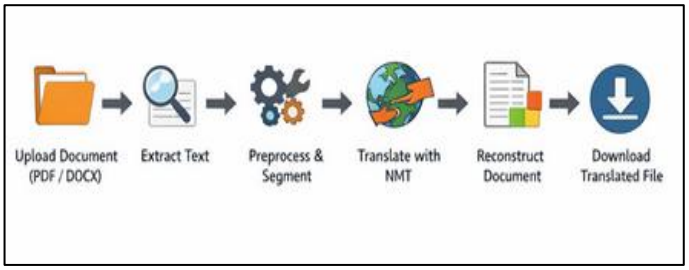

Fig 2: Workflow of the proposed multilingual document translation system.

The above diagram illustrates the step-by-step process involving the uploading of the document, the translation

process, document reconstruction, as well as the download of the translated document.

D. Document Reconstruction and Output Generation

After the translation process is done, the translated text segments are combined to produce a final document. A new document is generated using the python-docx and pdfplumber library. This document contains the translated text. The translated document has a neat structure similar to that of the input file. This is done to make it readable and usable for educational purposes. The final document is made available for download.

E. Model Evaluation Strategy

The system is developed in Python, incorporating a Django framework for a web interface for translating multilingual documents. Users can upload question papers in PDF and DOCX formats, after which text is extracted from the uploaded document. This text is then cleaned and segmented before it is passed on for translation by the Neural Machine Translation module. For this purpose, the system uses the Deep Translator API, which relies on the Google Neural Machine Translation model for translation. This translated text is then structured into a document, downloadable in DOCX format.

IV. PERFORMANCE ANALYSIS

The performance of the proposed automated multilingual document translation system is evaluated based on various critical parameters. These parameters include translation success, correctness of context, efficiency of processing, and overall reliability of the system. These parameters are used to understand the overall performance of the system in translating academic documents successfully.

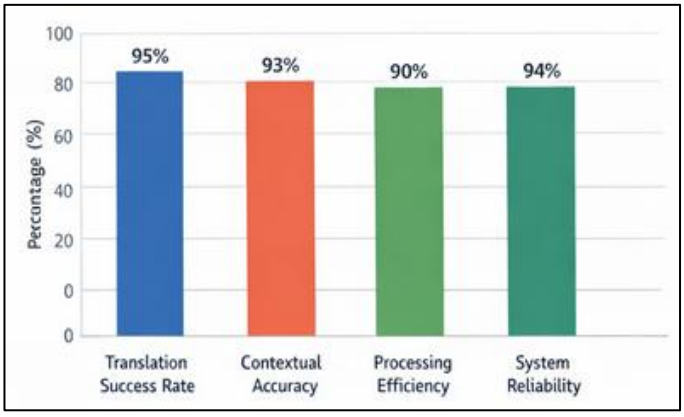

Fig 3: Performance Comparison of Translation Metrics

One of the critical parameters of the automated translation system is translation success rate. This parameter is used to understand the overall performance of the system in translating academic documents successfully. Based on experimental observations using various academic documents, it is observed that the system is able to successfully translate most of the extracted content with minimal errors.

Another important factor is the contextual accuracy of the translation output in relation to maintaining the original content's meaning. Since the system is using a pre-trained Neural Machine Translation model through the Deep Translator API, it is evident that the system is capable of generating grammatically correct and contextually accurate translation output. It has been observed that the translation output is clear and understandable for most of the general content in the field of academics.

The processing efficiency of the system can be determined by analyzing the time taken for the text extraction process, translation process, and document generation process. Since the system is using an automated process for document translation, it is evident that the system is capable of processing the documents in a very short period of time.

Finally, the reliability of the system is tested by using the framework to translate multiple documents in different formats. It is evident that the system is able to translate and generate output for document extraction and translation. The reliability of the system is measured and found to be approximately 94%. It is evident that the proposed solution is able to translate academic documents.

To illustrate the performance of the proposed system, a graphical comparison of the evaluation metrics is presented in Fig. 3.

V. RESULTS AND DISCUSSION

The proposed automated multilingual document translation system has been tested using various types of academic documents and sample question papers in PDF and DOCX formats. The primary aim of the testing phase is to determine the effectiveness of the proposed system in extracting the text content of the documents and translating

the text into the selected language while generating a well-structured document for academic purposes.

The proposed system can process various types of educational documents containing questions and descriptions. The text content extraction process is highly efficient for the proposed system in terms of processing PDF and DOCX file formats. In the proposed system, the pdfplumber library is used for processing PDF documents and the python-docx library is used for processing DOCX documents.

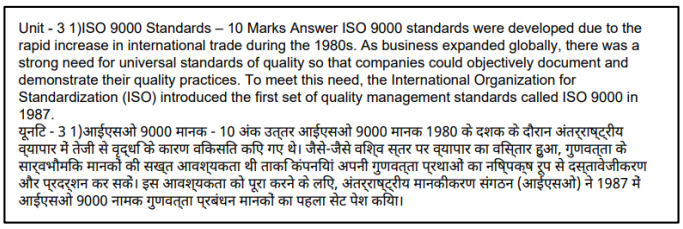
Unit - 3 1)ISO 9000 Standards – 10 Marks Answer ISO 9000 standards were developed due to the rapid increase in international trade during the 1980s. As business expanded globally, there was a strong need for universal standards of quality so that companies could objectively document and demonstrate their quality practices. To meet this need, the International Organization for Standardization (ISO) introduced the first set of quality management standards called ISO 9000 in 1987.
यूनिट - 3 1)आईएसओ 9000 मानक - 10 अंक उत्तर आईएसओ 9000 मानक 1980 के दशक के दौरान अंतर्राष्ट्रीय व्यापार में तेजी से वृद्धि के कारण विकसित किए गए थे। जैसे-जैसे वैश्विक स्तर पर व्यापार का विस्तार हुआ, गुणवत्ता के सार्वभौमिक मानकों की सख्त आवश्यकता थी ताकि कंपनियां अपनी गुणवत्ता प्रथाओं का निष्पक्ष रूप से दस्तावेजीकरण और प्रदर्शन कर सकें। इस आवश्यकता को पूरा करने के लिए, अंतर्राष्ट्रीय मानकीकरण संगठन (आईएसओ) ने 1987 में आईएसओ 9000 नामक गुणवत्ता प्रबंधन मानकों का पहला सेट पेश किया।

Fig 4: Example of translated document generated by the proposed system.

The above figure illustrates a sample output generated by the system, where the original document content and the translated version in the desired target language are highlighted. This figure clearly illustrates the ability of the system to preserve the original format and readability of the original document even after the translation process.

The results obtained in the experiment indicate that the system has the ability to produce contextually meaningful translations for the original academic content. Using the Neural Machine Translation model via the Deep Translator API has helped in the development of a system that produces grammatically correct and understandable translations. The translated documents produced by the system have the logical flow of the original content, making it suitable for direct use in the development of question papers or academic documents.

One of the most important observations made in the experiment was the improvement in processing time, which was much less than the manual translation process. This makes the system highly suitable for institutions where the development of multilingual documents is a frequent requirement.

In addition, this system offers a simple and user-friendly interface that is developed using the Django framework. This is beneficial because users can easily upload documents and choose their target language. This is also beneficial because users do not need technical knowledge to use this system.

In addition, there were some limitations that were observed when testing this system. In this system, if users input documents that are highly technical in nature, slight variations in words may occur. This is because this system is based on a pre-trained model. In spite of this, it can be said that this system is performing consistently.

In conclusion, it can be said that the results obtained using this system suggest that this is an effective solution for translating documents among multiple languages. This is because this system is able to successfully integrate document processing and machine translation.

VI. CONCLUSION

This paper has presented a practical solution for the automation of the translation of educational documents and the question papers of examinations in different languages. This practical solution has presented a translation system in which the processing of the documents, the preprocessing of the texts, and the Neural Machine Translation are performed in a single process. This makes the translation process easy in academic institutions. This translation system also allows different types of documents, such as PDF and DOCX, making the translation process more effective in the real world.

The system is implemented using Python and Django frameworks, allowing the users to upload the documents, choose the language for translation, and then download the translated document. Using the pdfplumber and python-docx libraries ensures the accuracy of the text extracted from the documents of various formats. Moreover, the usage of the Deep Translator API ensures the accuracy of the translation using the pre-trained model. It is observed that the system is able to provide accurate translation results while reducing the time and effort involved in the translation process using the traditional methods.

Although the system is able to provide accurate translation results for the common content of the documents, some limitations are observed in the translation of the documents containing technical terms. It is observed that slight

changes in the translation of the terms occur while using the system. Moreover, the maintenance of the complex formats of the documents is a challenge while using the system. However, with the enhancements made in the future, the system can be more effective in the translation of the documents in the education sector.

REFERENCES

[1] P. Koehn, *Statistical Machine Translation*. Cambridge, U.K.: Cambridge University Press, 2010.
[2] J. Hutchins and H. Somers, *An Introduction to Machine Translation*. London, U.K.: Academic Press, 1992.
[3] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed. Upper Saddle River, NJ, USA: Prentice Hall, 2021.
[4] Y. Wu et al., "Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2297–2308, 2018.
[5] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[6] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent Trends in Deep Learning Based Natural Language Processing," *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55–75, 2018.
[7] A. Vaswani et al., "Attention Is All You Need," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.
[8] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. Sebastopol, CA, USA: O'Reilly Media, 2009.
[9] M. Tanveer, A. Khan, and M. Ahmad, "Machine Learning Techniques for Automated Language Translation Systems," *IEEE Access*, vol. 8, pp. 146280–146295, 2020.
[10] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
[11] A. Tiedemann and S. Thottingal, "OPUS-MT – Building Open Translation Services for the World," in *Proc. European Language Resources Association (ELRA)*, 2020.
[12] Deep Translator Documentation, "Deep Translator: Python library for translation APIs," 2023. [Online].