WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

TranslateX

Venugopal YR[1], Pragya Kumari[2], Siddharth Pratim Kumar[3], Praveen Kumar K[4], Sujith Kumaar SK[5], Prof. Rahul Pawar[6]

[1,2,3,4,5]Student of MCA, Department of CS & IT, Jain (Deemed-to-be) University, Bangalore, India
[6]Assistant Professor, Department of CS & IT, Jain (Deemed-to-be) University, Bangalore, India
[1]venugopalyr.vs18@gmail.com, [2] pragyakumari10199@gmail.com, [3]siddharth77q@gmail.com, [4]praveen25kumar.k@gmail.com, [5]sujith972001@gmail.com, [6]rahul.pawar@jainuniversity.ac.in
DOI: https://doi.org/10.55248/gengpi.5.0324.0814

ABSTRACT:

The practice of translating voice or writing from one language to another while preserving the original content's meaning is known as language translation. The internet's growth and the growing globalization of commerce have made language translation a vital instrument for cross-cultural understanding and communication.

The first method of translating across languages was manual translation, in which a speaker of both languages would interpret voice or writing. But as technology has developed, machine translation has grown increasingly widespread. Machine translation is the process of automatically translating text or voice from one language to another using computer algorithms.

Rule-based machine translation, statistical machine translation, and neural machine translation are some of the methods used in machine translation. Creating a set of rules that control how words and phrases are translated from one language to another is known as rule-based machine translation. Statistical machine translation determines the most likely translation for a given input by using statistical models. Through the analysis of vast volumes of data, neural machine translation employs deep learning algorithms to learn how to translate languages.

Even with major advances in machine translation technology, translating language properly remains difficult, especially when dealing with idiomatic idioms and context-specific subtleties. Because accuracy is so important in some applications, including legal or medical translation, human translators are still required.

1. Introduction:

Language has been a major obstacle to communication for ages, and people have long attempted to find solutions for language translation problems. Humans have created a variety of methods for translating languages throughout the years in an effort to address the issues brought on by linguistic disparities. The real world is full of labels, important messages, and helpful information, but the majority of it is written in one of many official languages depending on the country in question. In addition, if a visitor does not speak the language of the nation they are visiting, it will be difficult for them to do their job there. To understand the message, they must utilize an internet translation service or keep a pocket dictionary on them. The introduction of optical character recognition (OCR) has made the process of digitalization easier for consumers. OCR, however, is unable to convert scanned text pictures into other languages that are legible by humans. In order to enhance text recognition, this article offered an Android-based application programming interface (API) that uses Firebase to convert scanned text images into the user's preferred language. The suggested API scans text pictures, extracts text from them, and converts the text into the user's selected language.

A language-to-language translation system is a technological advancement that facilitates automatic translation between two languages so that speakers of different native tongues may communicate with one another. Language to Language Translation Technology is used to translate a speech in one language to another voice in a different language. Speech Synthesis Technology translates text in one language into another; Speech Recognition Technology recognizes a person's utterance and turns it into a text; and Speech Synthesis Technology is responsible for turning the text into a speech. Furthermore, the Language to Language Translation System depends heavily on both natural language understanding technology and user interface technology that is connected with the UI (User Interface). As of right now, a device called Language Translation Technology may be purchased that translates free-form multilingual discussions quickly. Systems for translating language translate voice in real time. In order to provide top-notch translation for all users, challenges in completing the interpretation include overcoming speaker-dependent variables, such as differences in speaking style or pronunciation. Furthermore, in the real-world use of language translation systems, voice recognition systems need to be prepared to address external elements such as ambient noise or speech from other speakers. The issue of cross-lingual intent conversion with regard to speech intonation exists in the current system. Along with database work, the current system produced a parallel voice database for the English-Portuguese language pair. The current system offered word focus analysis for two language pairs, suggested an automated method for transforming accents into other languages, and objectively

demonstrated how the suggested methods improved TTS (Text to Speech) intonation contours. The translation method proposed in this work converts the language to text first, then to the target language, and finally to the target language using a dictionary. Automated speech recognition, or ASR, converts spoken phrases in a source language into text in that language. Machine translation then translates the source language text next to text in the target language. Lastly, a speech synthesizer converts target language text to speech.

2. LITERATURE SURVEY

Translation across languages is a crucial component of communication and has long been a research area. A translation management system, formerly known as a globalization management system, is a kind of software designed to maximize translator productivity by automating several steps in the human language translation process.

Up until the end of the 1980s, the predominant framework for MT research was based on several kinds of linguistic rules, such as morphological, syntactic analysis, lexical, lexical transfer, and syntactic generation rules. The Rule-based approach was the cornerstone of all Interlingua systems, both those that were primarily knowledge-based and those that were linguistics-oriented. It was most evident in the three leading transfer systems (Eurotra, SUSY, and Adriane).

However, since 1989, new techniques and approaches that are currently colloquially referred to as "corpus-based methods" have emerged, upending the rule-based approach's supremacy. First, in 1988, an IBM team released the findings of tests conducted on a system using just statistical techniques. Many researchers were surprised by the method's efficiency, which encouraged others to try out different statistical techniques in the years that followed.

Second, at the same time, several Japanese organizations started to publish preliminary findings utilizing techniques based on translation example corpora, or what is currently known as "example based translation." The main characteristic of both methods is that no syntactic or semantic rules are applied while analyzing texts or choosing lexical equivalents.

Key conclusions from reviews of the literature on language translation are as follows:

1. In addition to translating words, language translation also takes into account the subtleties and cultural background of the original language. As a result, translation is a difficult undertaking that calls for a thorough knowledge of both the source and destination languages.

2. The translation industry has seen a change thanks to the usage of technology. Systems for machine translation (MT) have been created to translate text mechanically between different languages.

3. Despite improvements in speed and efficiency, machine translation (MT) cannot match human accuracy in translation because of linguistic complexities and cultural quirks.

4. Rule-based, statistical, and neural machine translation are some of the methods used in language translation. The newest method, neural machine translation (NMT), has demonstrated encouraging outcomes in terms of translating text with greater accuracy.

5. A number of measures, including TER (Translation Error Rate), METEOR (Metric for Evaluation of Translation with Explicit ORdering), and BLEU (Bilingual Evaluation Understudy), can be used to assess the quality of language translation. These metrics assess translation quality by taking into account variables including sufficiency, correctness, and fluency.

6. The rise in international trade and globalization have led to a notable increase in the demand for language translation services in recent years. As a result, specialist translation services in domains including technical, legal, and medical translation have emerged.

7. It's conceivable that technological advancements like machine learning and artificial intelligence will drive language translation in the future. These technological advancements have the ability to increase language translation's efficiency and accuracy while also speeding up accessibility for users worldwide.

2.1SMT for Indian Languages:

(2010) Sanjay Kumar et al. released an extensive analysis of machine translation programs created for Indian languages. The majority of the systems (Mantra machine translation system, MaTra system, AnglaBharti Technologies, Anuvadak machine translation) are designed to translate text from English to Hindi.

Many machine translation systems created for both Indian and non-Indian languages were covered by Goyal V. et al. It is discovered that a few Indian systems employ the statistical approach in part. Statistical language models are used by Angla Bharti-II (2004) for automated post editing. Shakti (2004) uses a statistical technique in addition to a rule-based strategy for translating texts from English into Telugu, Hindi, and Marathi. IBM has begun development of a statistically based English-Hindi translation system. In order to translate English into South Dravidian languages like Malayalam and Kannada, a machine translation system employing statistical techniques was developed, as described by Unnikrishhanan et al. (2010). The aforementioned system was developed using a variety of tools, including MOSES decoder, which translates English to Malayalam (or Kannada), GIZA++, which trains translation models, and SRILM, which creates language models. The Stanford statistical parser, the converters from Roman to Unicode and Unicode to Roman, the morphological analyzer and generators, the English morphological analyzer, the morphological analyzers for Malayalam and Kannada, the morphological generators for Malayalam and Kannada, and the transfer rule file are additional tools utilized at different stages of the translation process.

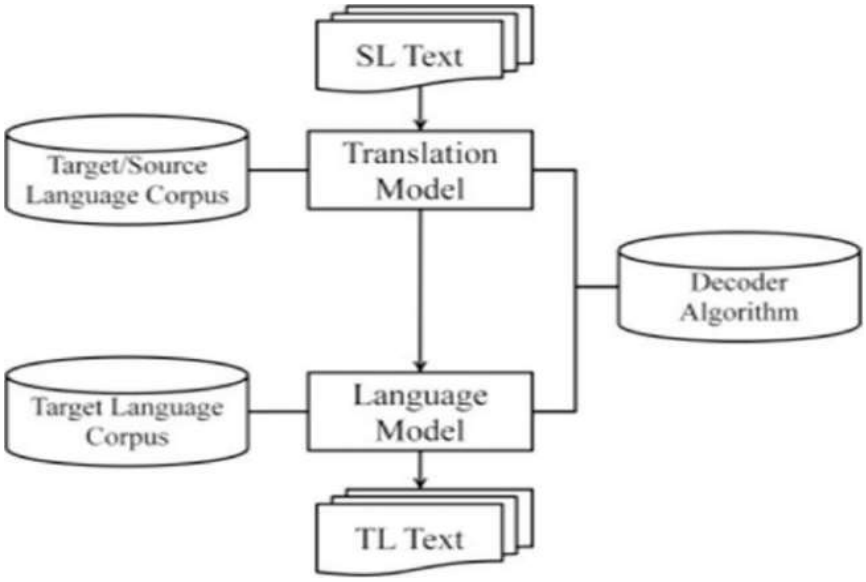


Fig:1 Architecture of SMT System

2.2 Natural Language Processing:

One area of machine learning called natural language processing (NLP) facilitates effective communication between computers and people. Recent years have seen tremendous advancements in the ability of computers to understand human languages through the use of human language. The main challenge is information overload, which makes it difficult to extract a specific, pertinent piece of information from enormous databases. Semantic and meaning awareness is crucial and challenging for summary systems because of issues with consistency and usability. The importance of the relationship between entities and objects must also be explained, particularly when using high-dimensional, diverse, complicated, and low-quality data, as is the case with other languages. For user requests in plain language to be fulfilled by RDBs, the requests must be translated into formal database queries such as SQL. Furthermore, in reality, NLP queries can be challenging to transform into structured database requests or service request URLs because of things like intricate DBLayouts with constraints, table names, and columns, or the semantic disparity between user-generated terms and DB-Nomenclature as it relates to languages other than English.

3.METHODOLOGY:

Two goals form the foundation of the proposed Android Text translation: (i) Word Transmission and (ii) Script Transfer Mark. Using MATLAB, text libraries of many languages were created for a mobile translator. These libraries contained the most common phrases that are used in everyday discussions. To convert to text production, the user is prompted to enter text in the Android code and click the button corresponding to their language.illustrates the mobile translator's system design.

Comprehending the Original Text: To fully understand the content, context, and subtleties of the original material, the translator either reads it or listens to it. It is essential to comprehend the idioms and cultural context.

Research and Clarification: To guarantee a correct translation, the translator looks up any unclear phrases, cultural allusions, or technical jargon in the original text.

Selecting a Translation Method: The translator chooses whether to interpret anything literally (word for word) or in a way that is more idiomatic or culturally relevant (dynamic or free translation), depending on the context.

Translating: After transforming the text from the source language to the target language, the translator starts the actual translation process. This might entail rearranging phrases, dissecting intricate sentences, or identifying synonyms.

Preserving Tone and Style: To ensure coherence and authenticity, the translator attempts to preserve the tone, style, and register of the original text in the target language.

Editing and proofreading: The translator checks the material for correctness, fluency, and clarity after completing the first translation. Proofreading assists in identifying any mistakes or discrepancies.

Cultural Adaptation: Translation is often influenced by cultural subtleties. To make cultural allusions, idioms, and humor clear and relevant to the intended audience, the translator modifies them.

Localization (if applicable): To accommodate language and cultural variations, more modification may be required if the translation is meant for a particular area or locality.

Quality Assurance: To guarantee accuracy and consistency, translations are frequently subjected to quality assurance reviews by additional translators or editors in professional settings.

Revisions and Comments: In order to enhance clarity and faithfulness to the source material, any required changes are performed in addition to incorporating comments from clients or reviewers into the translation.

Finalization: The translation is created, supplied to the customer, or published as required once all necessary adjustments have been made and it satisfies the required quality standards.

Continuous Improvement: To improve their proficiency and keep current with linguistic and cultural changes, translators constantly hone their abilities via practice, feedback, and further professional development.

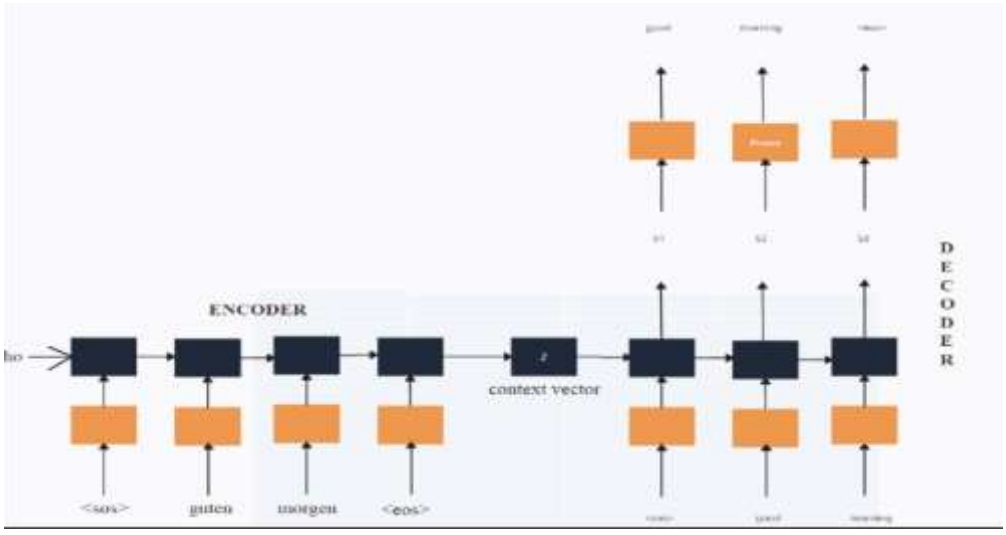
4.DESIGN DIAGRAM:

Sequence - to - Sequence Diagram:

In natural language processing (NLP), a sequence-to-sequence (Seq2Seq) diagram usually shows the architecture of a neural network model created for NLP tasks where the input and output are text sequences. An encoder and a decoder are the two primary parts of a typical Seq2Seq model. Here's how it's visualized in an NLP context:

- **Source Text (Encoder Input):** This is a representation of the input text, which may be a lengthy paragraph or document that you wish to condense. The Seq2Seq model receives this text as input.
- **Encoder:** The part of the neural network that processes the source text is called the encoder. It takes in the text input and outputs an encoded version of it. Capturing the most important information from the original text is the encoder's job.
- **Context Vector:** The source text's encoded data is represented by the context vector. It condenses the main ideas and information from the input text, which is what will be utilized to create the summary.
- **Decoder Input (Start Token):** The context vector is sent into the Seq2Seq model at the decoder input, where it begins to generate the output sequence—in this case, the summary. To mark the start of the output sequence, the decoder frequently starts with a specific token, such as the start-of-sequence token (<sos>).
- **Decoder:** Another part of a neural network is the decoder. The source text is summarized in the output sequence, which is produced using the context vector. Using the context vector and previously created components as input, the decoder constructs words or tokens one at a time.
- **Summary (Decoder Output):** This is the representation of the output summary. One word or token at a time is produced by the Seq2Seq model, and these components are then combined to create the source text's summary.

Figure :2 illustrates the Sequence-to-Sequence diagram of ”Translate X” project.



5.IMPLEMENTATION:

DESCRIPTION:

This source code appears to be an implementation of a PyTorch and TorchText-based sequence-to-sequence (Seq2Seq) machine translation paradigm. This is a detailed explanation of the code:

Import Necessary Libraries:

- The necessary libraries, such as PyTorch, TorchText, spaCy, and random, are first imported into the code.

Data Preprocessing:

● Two functions, tokenize_ger and tokenize_eng, are defined in the code to tokenize text in German and English using the spaCy tokenizers.

● It initializes the two TorchText Fields—english and german—that will be utilized for vocabulary development and tokenization.

Load and Process Datasets:

● Using the Multi30k dataset from TorchText, the code defines the source and target language extensions ('.de' for German and '.en' for English) along with the previously defined Fields.

● Test, validation, and training sets make up the dataset.

Vocabulary Building:

- The training data is utilized to construct vocabulary in the German and English fields. Words must appear in a vocabulary at least twice in order to be included, and the word count is limited to 10,000.

Encoder Implementation:

The 'Encoder' class is defined, which is a part of the Seq2Seq model. The encoder takes the following parameters:

- **input_size:** Size of the input vocabulary.
- **embedding_size:** Size of the word embeddings.
- **hidden_size:** Size of the encoder's hidden states.
- **num_layers:** Number of LSTM layers in the encoder.
- **p:** Dropout probability.

The encoder consists of an embedding layer and an LSTM layer.

Decoder Implementation:

- The Decoder class is defined, which is the other part of the Seq2Seq model.
- The decoder takes similar parameters as the encoder, but also requires an output_size, which is the size of the target vocabulary.
- The decoder consists of an embedding layer, an LSTM layer, and a linear layer for generating predictions.

Seq2Seq Model:

- The Seq2Seq class is defined, which combines the encoder and decoder.
- The model takes the encoder and decoder as input.
- In the forward method, it performs the following steps:

I. Accepts source and target sequences, as well as a teacher_force_ratio (used during training for teacher forcing).

II. Initializes tensors for storing the output predictions.

III. Passes the source sequence through the encoder, obtainind the final hidden and cell states.

IV. Initializes the decoder input with the SOS token from the target sequence.

V. Iterates through the target sequence one step at a time:

Machine translation and other sequence-to-sequence activities are the target audience for this Seq2Seq model. The input sequence is processed by the encoder, and the output sequence is produced step-by-step by the decoder, which makes predictions along the way. The model may be applied to challenges involving machine translation for both training and inference.

CONCLUSION:

Our language translation platform is dedicated to providing top-notch services that meet the various demands of people and organizations worldwide. We recognize the value of efficient communication in the globalized world of today, when language boundaries can obstruct chances for understanding, cooperation, and growth.

We guarantee accuracy in every translation assignment we take on thanks to our staff of skilled translators and state-of-the-art equipment. Our translators can faithfully capture the subtleties and context of the original material since they are not just multilingual but also have subject-matter experience. We check and polish our translations using strict quality control procedures to make sure the finished product is up to par.

Since different languages have different meanings and structures, great care must be taken when translating written or spoken texts to ensure that the original meaning and sentence structure are preserved. After all, the goal of any translation is to convey the original meaning of the target language context.

REFERENCES:

[1] How a Typical Translation Project Works from Start to Finish.https://www.linkedin.com/pulse/how-typical-translation-project-works-from-start-finish- uli-dendy.

[2] Munday, J. (2016). Introducing Translation Studies: Theories and Applications (4th ed.). Routledge.

[3] Luong, Minh-Thang, Pham, Hieu, and Manning, Christopher D. (2015). Effective Approaches to Attention-based Neural Machine Translation. Conference on Empirical Methods in Natural Language Processing.

[4] Papineni, Kishore, Roukos, Salim, Ward, Todd, and Zhu, Wei-Jing. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. Association for Computational Linguistics.

[5] Hieber, Felix, Domhan, Tobias, Denkowski, Michael, and Vilar, David. (2017). Sockeye: A Toolkit for Neural Machine Translation. Conference on Empirical Methods in Natural Language Processing.