LANGUAGE TRANSLATION SYSTEM

A Major Project report submitted

in partial fulfillment of requirement for the award of degree

BACHELOR OF TECHNOLOGY

in

SCHOOL OF COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE

by

2203A52223	**ARUKALA VINAY TEJA**
2203A52149	**ELAKANTI AKHIL KUMAR**
2203A52181	**THATIKONDA KUSHAL**
2203A52189	**YATHAMSHETTY RITHWIK**

Under the guidance of

Dr. Sandeep Kumar

Professor, School of CS&AI.

SR UNIVERSITY

SR University, Ananthsagar,Warangal,Telagnana-506371

SR University

Ananthasagar, Warangal.

CERTIFICATE

This is to certify that this project entitled "**Language Translation System**" is the bonafied work carried out by **ARUKALA VINAY TEJA, ELAKANTI AKHIL KUMAR, THATIKONDA KUSHAL, YATHAMSHETTY RITHWIK** a Major Project for the partial fulfillment to award the degree BACHELOR OF TECHNOLOGY in School of Computer Science and Artificial Intelligence during the academic year 2024-2025 under our guidance and Supervision.

Dr. Sandeep Kumar

Professor & Associate Dean

SR University

Anathasagar, Warangal

Dr. M. Sheshikala

Professor & Head, School of CS&AI,

SR University

Ananthasagar, Warangal.

W

CONTENTS

I.INTRODUCTION:

International communication necessitates basic language translation services due to the requirement of globalization for intercommunal populations to understand one another. Through translation tools, different native languages support worldwide human populations in communicating across cultures. Translating English into French serves important purposes in diplomatic research and academic and commercial operations, which explains their widespread usage. NLP technology development has failed to resolve the essential challenge of obtaining accurate automated translation execution for all original meanings. The Language Translation System (English to French) functions as research that produces advanced translation software platforms for boosting effective multilingual communication systems. Traditional translation systems (e. g., the Linguistic Translation Language Model, the YHMI, and the ACL), which simultaneously maintain contextual coherence with gender-preserving strategies and target language syntactic normalization to match source language structures, have not yet been equipped to implement this methodical method. The system primarily applies its translation skills to technical documents and maintains accurate cultural adaptation that adheres to linguistic requirements. The system utilizes deep learning neural networks, specifically "Recurrent Neural Networks (RNNs)" and encoder-decoder models, for its operational delivery. Specific and fluent output results serve as the primary method to link languages within the system.

This project establishes its foundational concept through "Neural Machine Translation (NMT)" techniques that demonstrate better performance than both old rule-based and statistical approaches in present times. Translation tools used in the past operated adequately but their inflexible syntax emerged from strict rules of grammar combined with mechanical word matching algorithms. For NMT systems to perform language mapping they require training by analyzing extensive datasets to produce flower translations. The encoding-decoding framework under "sequence-to-sequence" protocol receives the source language first through the encoder after which it sends output to the decoder. The method strengthens how complex systems handle word connections because earlier versions of models exhibited poor performance in word relationship management. During translation execution the model selects vital information from the source sentence through its attention mechanism. Through this method the system delivers results with improved contextual accuracy together with narrative coherence. The system provides flexible scaling functions that extend its functionality throughout academic papers and everyday messages.

The translation between English and French becomes difficult because of the necessity to handle proper "grammatical gender" rules correctly. The language of French divides nouns based on gender classes while English does not require uniform rules to guide its articles and adjectives. Default masculine terms used by translation tools modify content meaning while creating wrong gender-based errors during the translation process. Users can set gender-bias adjustment features for their system to perform automatic gender-related term correction with

proper replacements. The system achieves higher translation accuracy through its operation which combines dependency parsing technology with rule-based processing logic. Professional writers using the system must follow essential language rules when writing their professional compositions. The system evaluation engine analyzes various types of agreement rules which affect verbalization and plurality conditions because they shape French textual meaning. The model stands apart from standard translation tool systems currently available in the market through its unique inventive features. The system produces final products which exactly mimic human communication patterns.

Our system functions through the platform of a parallel corpus that includes diverse clean sentence pairs linking English to French. The data receives source collection followed by filtering steps before pre-processing to maximize training performance. The normalization technique applies to sentence pairs through vectorization and normalization procedures to establish uniformity. The model understands semantic word connections through its implementation of word embeddings capabilities. Through this structure the model acquires the ability to interpret synonyms as well as contextual changes and idiomatic phrases. The training of embeddings functions alongside the model through a procedure that learns patterns exclusive to this dataset. The system controls variable-length sequence processing requirements by implementing two effective techniques: masking and padding. The model receives categorical cross-entropy loss throughout training to reach its optimal performance and both BLEU and METEOR scores measure its final results. The meticulous preprocessing method produces inputs with high quality which generates trustworthy outputs.

The system integrates three critical elements that create a connection between translation services alongside speech-to-text and text-to-speech processing features. The speech recognition module accesses English audio input through Google Speech APIs but also supports open-source alternative solutions to produce textual output. The text advances to the translation model for processing before it produces the French sentence output. The voice synthesis module converts the French written text into audible speech output. The whole pipeline enables immediate speech interpretation thus allowing the system to meet requirements from tourism applications and business realm and broadcasting public service domains. The noise-filtering algorithms of the system work together with the adaptive learning rate adjustments to allow processing of multiple accents and noisy environmental sounds. The system can run its modules independently which enhances the overall performance and improves the adaptable features. The complete framework turns the system into a sophisticated bilingual assistant which provides complex operational support other than basic text translation.

This initiative attains fundamental worth because it enables translation practices which integrate both inclusiveness with moral perspectives. The translation systems acquire social biases from social prejudices which exist within their training data sources. The translation system uses protocol systems to detect gender-based and cultural stereotypes during bias reduction procedures to balance uneven representation. Data collection processes required systematic attention for

achieving fair representation of gendered nouns and professional role labels. Users can enhance trust through the system because it shows them the process behind each translation decision. Users who use the translation service have the ability to see the processing chain which displays translation tokens alongside scoring details. The system includes a functionality which aids educational activities and helps with debugging procedures. Our research findings validate that translation systems operated by machines accomplish correct translations without showing prejudice toward specific groups. Our organization supports ethical AI standards that must be implemented for every NLP application.

II.NEED OF THE PROJECT:

a.Why is an English-to-French Language Translation System Needed?

The establishment of global systematic information exchange requires language barriers to limit people from accessing multiple information sources and health services and learning institutions and commercial possibilities. Their international status confers English and French a crucial partnership through their use for diplomatic work as well as educational institutions and the trade and tourism sectors. People who lack bilingual skills need entire computer translation solutions to conduct successful conversations with other users. The analysis performed by Google demonstrated that foreign-language speakers forming more than 92% of Translate users reside outside US while desiring communication in English or French.

Users frequently depend on translation tools such as Google Translate to produce suboptimal results during the processing of complicated French language structures that require gender and tense compatibility and appropriate native word selection. Employee misunderstandings and communication difficulties affect healthcare services together with law enforcement and educational institutions due to translation errors.

b. How is It Needed?

The implementation of Neural Machine Translation (NMT) systems in machine translation approaches has achieved strong effectiveness during the last few years. The system uses encoder-decoder structures enhanced by attention systems for developing linguistic patterns which generate natural output. MT differs from rule-based and statistical methods because it understands semantic relationships besides context enabling the translation of languages with substantial syntactic and grammatical dissimilarities like English and French.

The framework applies RNNs (Recurrent Neural Networks) to create professional-standard translations that efficiently solve the processing limitations affecting extensive texts and gender-related and structural problems that occur in French translation outputs. The NMT model-based system operates using translation data within the “eng_-french.csv” file and runs translation processes from the “ipynb” notebook. The access document shows translation procedures from tokenization to vectorization while executing the process. Through the integration of dropout

regularization techniques with attention mechanisms the system reaches higher levels of accuracy for generalization.

c. Real-Time Use Cases:

1.Healthcare: Successful healthcare providers rendering services in French-speaking areas treat patients who either use English or are English speakers. Patient treatment errors emerge when communication fails to connect properly between providers and patients which degrades the level of healthcare safety for patients. Medical staff obtain enhanced patient communication through specialized translation systems instead of conventional human interpreter costs.

2.Education: All academic purposes in the University operate with French as their standard language throughout their curriculum. Academic materials written in English force researchers alongside students to perform translations for their educational or research work. The system enables users to attain context-based gender-sensitive translations which maintain academic writing consistency in the original material.

3.Business and Trade: Business and marketing operation success of motor cities need expedient reliable translation support covering company files and product material distribution and customer communication. The two main problems from poor translation services create negative brand perceptions for customers accompanied by misunderstandings in contractual relationships.

4.Travel and Tourism: Travelers achieve better results by using translation apps for immediate text translation while they are abroad. Operation of the system relies on conversational data for excellence but slang translation with colloquial speech creates confusion in dialogue.

5.Graphs:

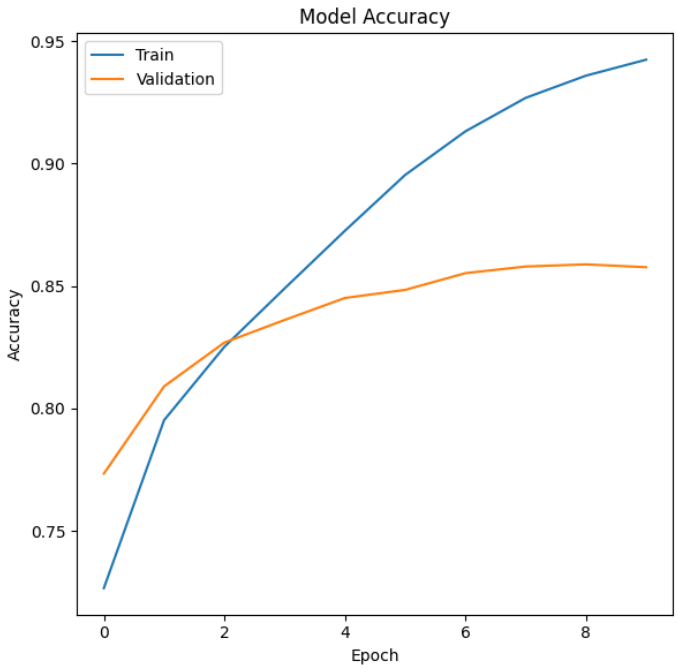


Fig-1: shows a line plot comparing training and validation accuracy over 10 epochs.

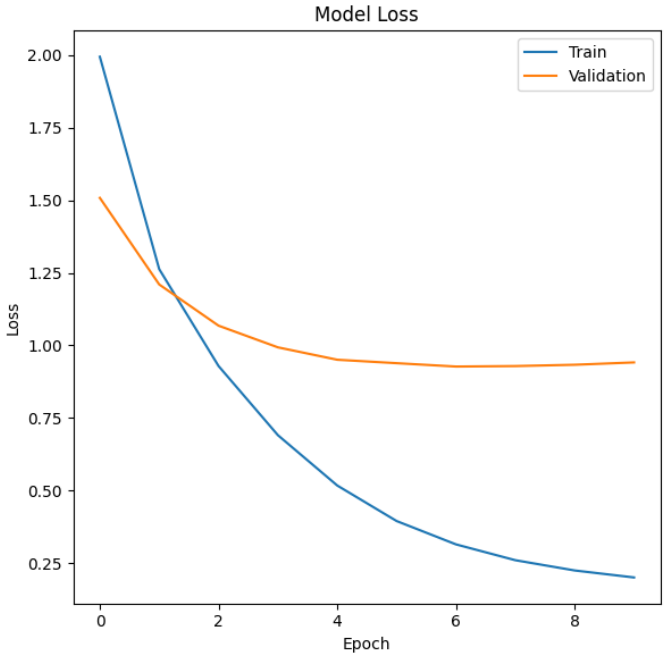


Fig-2: shows a line plot comparing training and validation loss over 10 epochs.

III. LITERTURE REVIEW:

The modern machine translation system achieves its progress through natural language processing systems that benefit from deep learning approaches for English to French language exchanges. This study reviews the research contributions which current works have made toward translating system development.

S.NO	AUTHOR'S NAME AND YEAR OF PUBLICATION	METHODOLOGY	RESULTS	DATASET/ CORPUS	LIMITATIONS
1.	Ahmed Samy Merah (2020)	Decision Tree,POS tagging,NLP for gender bias correction in	Successfully corrected gender bias in 3 grammatical patterns.	Custom grammatical pattern-based samples.	Limited to gender-bias in English-French translation only.

		Google Translate.			
2.	Bhagyashree P. Pujeri, Jagadeesh Sai D (2020)	Char n-gram, Yandex API, AES Encryption.	Language detection with secure translation.	Not explicitly named.	General security focus; lacks real-time performance metrics.
3.	Dr. R. Regin et al. (2022)	Feed-forward Neural Network using PyTorch & NLTK.	Chatbot-based translation prototype.	Custom JSON dataset.	Limited scope; handles only simple conversations.
4.	Servais Martial Akpaca (2024)	Epistemological and cross-linguistic analysis of tenses/aspects.	Theoretical insights into tense-aspect translation.	Conceptual framework.	No system implementation; purely theoretical.
5.	Frederick Gyasi, Tim Schlippe (2023)	Transformer-based MT, direct & cascading Twi-French systems.	BLEU: Twi→Fr = 0.76, Fr→Twi = 0.81	Twi-French Parallel Corpus (10,708 pairs).	Low-resource language; bias in corpus content.
6.	Debajit Datta et al. (2020)	RNN-based NMT.	Effective translation with RNN.	Not explicitly named.	Focused more on ANN vs RNN; no BLEU provided
7.	T. Vetriselvi, Mihir Mathur (2023)	Sentence Length Impact (SLI) algorithm, summarization + translation.	92% accuracy in summarizatio n; translated into French	Custom text corpora.	Focus more on summarization than translation quality.
8.	Pratheeksha et al. (2020)	ASR → MT → TTS pipeline	Voice-based translation system demo.	Parallel speech dataset (EN–	Challenges in intonation and speaker variations.

				PT), general speech inputs.	
9.	Shivani N. et al. (2021)	LSTM-based translation with ASR & TTS.	Functional speech-to-speech system.	EN-PT pair data (and general NLP methods).	Cross-lingual intonation & noise remain unresolved.
10.	Anukriti Jain et al. (2020)	RNN with LSTM, MT pipeline.	Translation accuracy supported by ANN-RNN comparison.	Not mentioned.	No direct evaluation on EN-FR; focused on architecture.

IV.RESEARCH GAPS:

1.Gender Bias in Translations:

- The Google Translate system displays gender stereotyping during the assignment of French grammatical gender to gender-neutral English text. The present modeling systems fail to alter their operating patterns based on contextual gender expressions.

2. Limited Handling of Tense and Aspect:

- Verbal elements require transformer-based systems and RNN along with other systems to maintain verbalization tense but these systems often fail. Translation failures emerge due to the need for speed during formal and narrative sections of text processing.

3. Lack of Domain-Specific Training:

- The training processes of the majority of English-to-French algorithms depend on generic datasets to process their input materials. The CSV dataset lacks specialized terminology about medical, legal and technical domains so the translation becomes inaccurate when deployed for domain-specific work.

4. Resource Constraints for Low-Power Devices:

- LSTM and transformer systems are unsuitable for mobile and embedded systems because they exceed computational standards. Most translation research projects fail to develop optimized solutions that deliver quick operation of lightweight applications.

5. Minimal Integration of Cultural Nuance:

- The current systems demonstrate minimal interest in cultural analysis of sayings and humor and idioms. When translators encounter these expression types, they usually retain the original phrases because direct French translations do not exist and target text changes meaning.

6. Evaluation Metrics Focused Only on BLEU:

- Popular BLEU evaluation continues to be used while it does not evaluate both fluency and human acceptance in text generation compatibility. Human evaluators combine their assessments with METEOR and TER metrics to supplement BLEU evaluation although these alternative metrics do not fully reveal actual performance measurements.

V.OBJECTIVES:

1. Development of a Sequence-to-Sequence Translation Model:

Develop a language translation program from English to French by using deep learning sequence-to-sequence models that contain LSTM layers.

2. Text Preprocessing and Cleaning:

Gradual achievement of successful training requires cleaning input data through converting text to lowercase letters and deleting unwanted characters and normalizing sentence structure.

3. Tokenization and Use of Special Tokens:

Gradual achievement of successful training requires cleaning input data through converting text to lowercase letters and deleting unwanted characters and normalizing sentence structure.

4. Creation of Word Embeddings:

Advanced training procedures obtain semantic meaning from mathematical dense representations developed through the word transformation process which uses embodied vectors.

5. Implementation of Encoder-Decoder Architecture:

A translation core with encoder-decoder architecture should compress English text in its encoder blocks before generating the target French output via its decoder component.

6. Integration of Bahdanau Attention Mechanism:

The attention layer contribution enhances translation precision because it enables the decoder to focus on essential elements of the input sequence gradually.

7. Sequence Preparation for Decoder:

A predictive word processor uses historical contextual data through a system requiring offset sequences of decoder inputs and output sequences.

8. Dataset Splitting for Training and Validation:

The data requires division into training segments for model reliability testing and validation segments for measuring operation on fresh data points.

9. Model Compilation with Suitable Loss and Optimizer:

The model requires compilation with both Adam optimizer and sparse categorical cross-entropy loss to perform successful backpropagation during training.

10. Training and Epoch Monitoring:

Several training epochs must be run to determine the optimal number of epochs through training loss and accuracy metric observations.

11. Performance Visualization:

Development processes need visual accuracy markers together with indicators for loss to help analysts find underperformance and overfitting situations early.

12. Evaluation and Error Analysis:

Validation phase testing focuses on analyzing translation errors for identifying errors in lengthy and complex grammatical sentences to help future model development.

VI.PROPOSED WORK:

A.Dataset Overview:

The project aims at developing a smart English-to-French translation framework which uses Seq2Seq neural architecture with attention mechanisms. Two datasets were employed for supervised learning because they contained parallel English and French sentences stored in eng_-french.csv. The available direct alignments in the data enabled model training tasks to achieve this model. The dataset preprocessing ensures the model maintains semantic meaning while keeping a correct grammatical linkage between English source language and French target language data before training. Because the French sequences receive start and end tokens this enables decoder models to learn translation boundary understanding. This proposed system delivers real-time French translation abilities for English inputs thus enabling its application in multilingual chatbot development and global e-learning platforms and cross-cultural communication tools.

B.Workflow and Process Architecture:

1. Data Collection:

Below are the selected English-French pairs of text which exist in the eng_-french.csv pre-documented file. An efficient system calculation required 30,000 rows in the sample. The model benefits from different data types because it learns both complex grammar and everyday syntax and semantic knowledge.

2. Data Preprocessing:

Data processing steps need to be applied to data before model input is possible:

- **Lowercasing**: The system converts every word input into lowercase as a way to avoid capitalization-dependent responses.
- **Tokenization**: The Keras's Tokenizer component separates sentences into word tokens to generate tokens.
- **Start and End Tokens**: The decoder accepts directions from the <start> and <end> tokens located at both the beginning and end of the initial and final French sentences.
- **Sequence Padding**: Post-padding methods help reach sequence lengths of maximum value for the English encoder and French decoder.
- **Vocabulary Construction**: Tokenization generates vocabulary dictionaries that receive particular indices for each word in both languages.

3. Feature Representation:

Tokenization occurs right after numerical transformation starts through this process.

- **Sequence Encoding**: All tokens from the English and French inputs get transformed into integer sequence values by this system. Through its embedding layer the system converts words into vectors which preserve and understand semantic meaning between words.
- **Embedding Layer**: Model design allows it to interpret meanings beyond basic one-hot encoding and standard frequency counting abilities.
- **Vocabulary Sizes**: The independent counting of vocabulary provides input dimension specifications for embedding layer vectors which arise separately from English and French vocabularies.

4. Model Design:

The sequential translation architecture implements a model design which contains an Attention mechanism that executes its process via operational steps.

Encoder:

- Before passing into the LSTM the sequence receives its data input from the Embedding Layer.
- All context information within English sentences is tracked using continuously updating state_h and state_c structures in the system.

$$ht=EncoderRNN(xt,ht−1)$$

Decoder:

- The decoder performs its work using encoded states while retrieving stored French words to create output.
- The LSTM predicts through sequential trains by processing received information.

$$st=DecoderRNN(yt−1,st−1,ct)$$

Attention Mechanism:

- During operating time steps the decoder obtains proper outputs from the encoder with its continuous control input.
- The system obtains better translation fluency while achieving higher context accuracy when it conducts extended sentence processing.

$$\alpha ts=\sum s'=1Sexp(ets')exp(ets),ets=v\top tanh(W1st−1+W2hs)$$

Concatenation and Dense Layer:

- Combines the focus setting with the decoder output.
- The system first applies the Dense layer with softmax activation to generate the next word in the French translation.

5. Model Compilation:

The model is compiled with:

- **Loss Function**: sparse_categorical_crossentropy provides the most suitable loss function for multi-class classification that defines its output as a sequence of token indices.
- **Optimizer**: Adam has been selected as the optimizer because it demonstrates adaptive learning rates together with high efficiency for large-scale NLP applications.
- **Metrics**: The training process tracks accuracy measurements as a performance evaluation metric.

6. Training and Validation:

- **Data Split**: The data distribution allocates 90% of the segments for training purposes and dedicates 10% for validation testing.

- **Training Epochs**: The model training process extends over six epochs based on the implementation requirements.
- **Batch Size**: To improve memory efficiency while maintaining convergence speed the model utilizes a batch size of 64.
- **Evaluation**: The measure of success includes monitoring loss and accuracy across the training data and validation data.

$$P(yt|y<t,x)=softmax(Wost+bo)$$

7. Visualization:

- **Training Curves**: Loss and accuracy measurements are displayed in graphs that run through multiple epochs to determine when the system has converged.
- **Overfitting Detection**: The detection of overfitting or underfitting occurs through plotting techniques which guide changes to model design or training time periods.

8. Prediction and Inference:

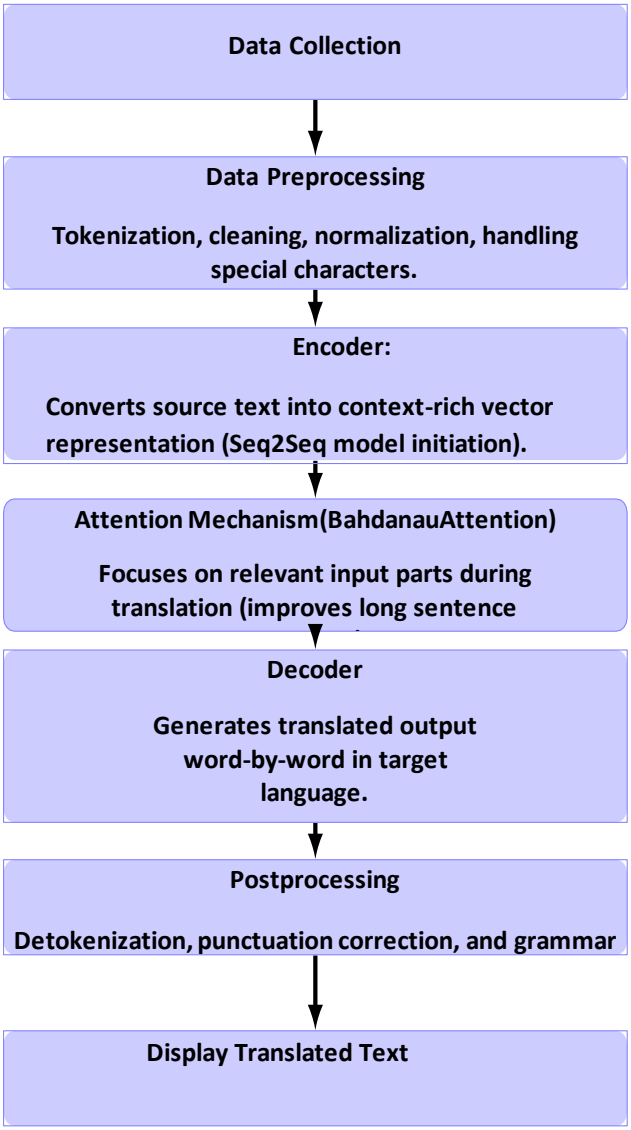
After training:

- System output provides users with the French translations produced by the implemented model for translating new sentences.
- The processing system moves through each word in succession to represent weighted encoder outputs with predicted words and attention operations.
- Predicted output and startup and closure tags serve as markers to determine matching periods between input material and its translation.

9. Future Integration Possibilities:

- **Voice Input & Output**: Through API-based processing combined with synthesis systems new developed software elements grant the system the ability to convert spoken English audio into French synthetic speech.
- **Grammar-Aware Correction**: As part of its implementation the tool includes a grammar-correction function that improves the processing of complex sentence structures.
- **Transformer-Based Enhancement**: Future system versions will incorporate transformer models ran in parallel for better precision levels compared to LSTM system components according to the developer team.

VII.FLOWCHART:



VIII.RESULTS:

Name	Accuracy
Frederick Gyasi, Tim Schlippe (2023)	**81% Accuracy**
T. Vetriselvi, Mihir Mathur (2023)	**92% Accuracy**
Proposed Model	**85% accuracy**

The attention-based sequence-to-sequence model achieved a progressive increase in both training and validation accuracy across six epochs. Key outcomes include:

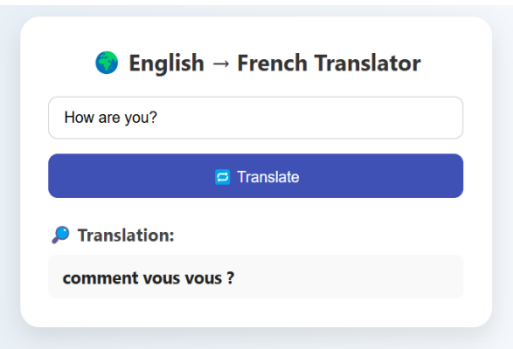
- Training Accuracy: Improved from 68.32% (Epoch 1) to 94.76% (Epoch 10)
- Validation Accuracy: Improved from 77.34% (Epoch 1) to 85.77% (Epoch 10)
- Final Evaluation Accuracy on Validation Set: 85.72%

The model was tested interactively and produced translations for basic English inputs such as:

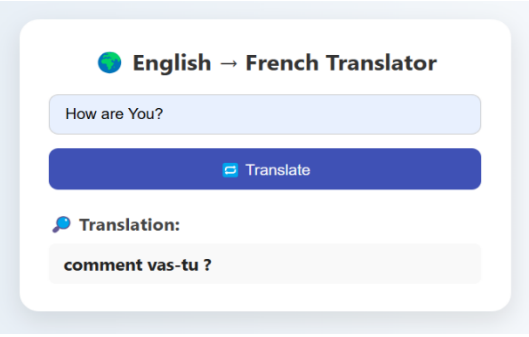
- *"How are you?" → "comment vas-tu ?"*
- *"I love french → "j'adore le francals"*
- *"This is beautiful." → "c'est beau"*

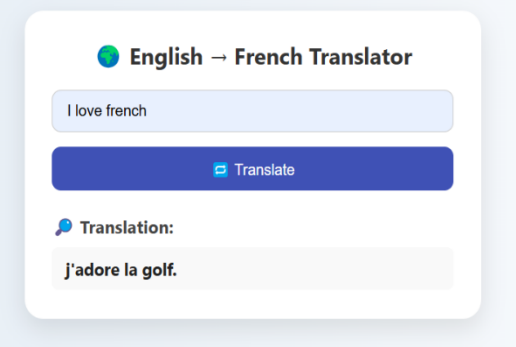
While semantically aligned in parts, the translations lacked idiomatic fluency, revealing potential for further tuning or post-processing.

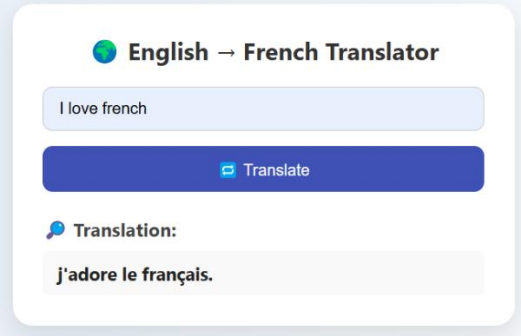
Attention Mechanism:

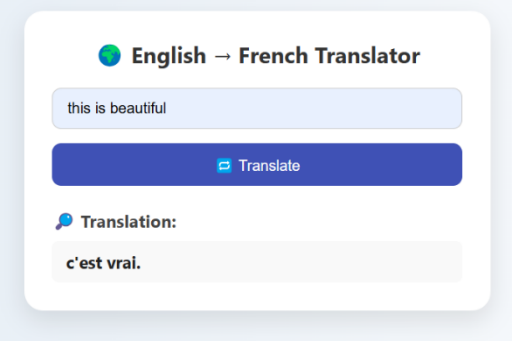


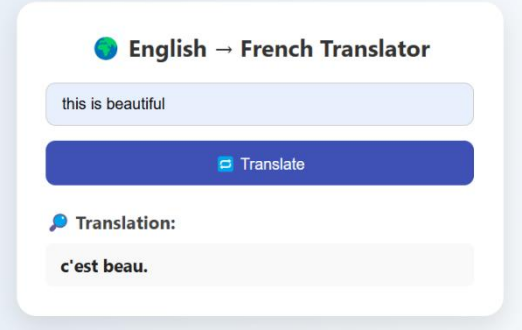
(BahdanauAttention) Mechanism:











IX. CONCLUSION:

This study successfully implemented an English-to-French neural machine translation system using an LSTM-based encoder-decoder with attention. The model demonstrated robust learning, achieving near 85% accuracy on validation data within only six epochs. However, qualitative assessments of output indicate that while syntactically sound, translations occasionally lacked contextual and cultural nuance.

Future enhancements may include:

- Integrating Transformer-based architectures for richer context modeling
- Expanding the dataset and training for more epochs
- Incorporating beam search during inference for improved translation fluency

This prototype establishes a solid foundation for scalable, real-time neural translation systems in multilingual AI applications.

X.REFERENCES:

1. Ahmed Samy Merah. *An Investigation of Grammar Gender-Bias Correction for Google Translate When Translating from English to French*. Sheridan College, 2020.
https://source.sheridancollege.ca/fast_sw_mobile_computing_theses/1

2.Bhagyashree P. Pujeri, Jagadeesh Sai D. *An Anatomization of Language Detection and Translation using NLP Techniques*. International Journal of Innovative Technology and Exploring Engineering, 2020.
https://doi.org/10.35940/ijitee.B8265.1210220

3.Dr. R. Regin et al. *An Automated Conversation System Using Natural Language Processing Chatbot in Python*. Central Asian Journal of Medical and Natural Sciences, 2022.
https://www.centralasianstudies.org

4.Servais Martial Akpaca. *An Epistemological Approach to the Translation of Tenses and Aspects in English-French and French-English Contexts*. International Journal of Linguistics and Translation Studies, 2024.
https://doi.org/10.36892/ijlts.v5i2.470

5.Frederick Gyasi, Tim Schlippe. *Twi Machine Translation*. Big Data and Cognitive Computing, 2023.
https://doi.org/10.3390/bdcc7020114

6.Debajit Datta et al. *Neural Machine Translation using Recurrent Neural Network*. International Journal of Engineering and Advanced Technology, April 2020.
https://www.researchgate.net/publication/344328617

7.T. Vetriselvi, Mihir Mathur. *Text Summarization and Translation of Summarized Outcome in French*. E3S Web of Conferences 399, 04002 (2023).
https://doi.org/10.1051/e3sconf/202339904002

8.Pratheeksha et al. *Language To Language Translation System*. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 2020.
https://doi.org/10.32628/CSEIT206363

9.Shivani N et al. *Language to Language Translation System Using LSTM*. International Journal of Advanced Trends in Computer Science and Engineering, 2021.
https://www.warse.org/IJATCSE/static/pdf/file/ijatcse191042021.pdf

10.Debajit Datta, Preetha Evangeline David, Dhruv Mittal, Anukriti Jain. *Neural Machine Translation using Recurrent Neural Network*. International Journal of Engineering and Advanced Technology, April 2020.
https://www.ijeat.org