International Journal of Applied Engineering & Technology

DEPLOYING A DATA SCIENCE MODEL TO BUILD A TN_DS_TRANS APP FOR TRANSLATING ALL THE TN GOVERNMENT FILES IN ENGLISH TO TAMIL

Dr. I. Priya Stella Mary[1]

[1]Assistant Professor, Department of Data Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India

ABSTRACT

This research paper aims to address the critical need for accurate translation of Tamil Nadu (TN) government files from English to Tamil, aligning with the government's efforts to promote Tamil as an administrative language. In this paper, a Data Science model using TensorFlow has been deployed to build a TN_DS_TRANS app. This app facilitates the translation of TN government files accurately, contributing to the government's mission of implementing Tamil as an administrative language at the district level. The proposed TN_DS_TRANS model achieved a BLEU score of 0.85, outperforming existing models such as Neural Sentence Rewriting (0.75), Data-Oriented Parsing Technique (0.79), and RNN Search Model (0.83), indicating its superior translation accuracy and efficiency.

Index Terms – Data Science, Machine Learning, Tamil Language, Translation.

INTRODUCTION

Language serves as a cornerstone of cultural identity, acting as a conduit for preserving heritage and fostering communication within a society. The Tamil language stands as an indomitable Classical Language, cherished as a profound treasure cultivated by humanity. Recognizing the unparalleled efficacy of one's mother tongue in uninhibited expression, the enactment of the Official Language Act underscores the efforts to establish Tamil as the language of instruction in Science and Technology. As the custodians of this linguistic legacy, the Government of Tamil Nadu has embarked on a commendable journey to establish Tamil as the administrative language across all government offices at the district level. Against the backdrop of this linguistic evolution, Data Science Technology emerges as a transformative force in the domain of translation. Harnessing machine learning and neural networks, translation apps have evolved into robust tools, ensuring precision and ease for translators.

One of the formidable challenges in this linguistic transition is the accurate translation of government files from English to Tamil, ensuring that the essence and details are faithfully preserved. Recognizing this imperative, the proposed research work, a Data Science Model based TN_DS_TRANS app has been created to harness the power of Data Science to create an innovative solution to facilitate seamless translation. To propel Tamil into the modern world, it becomes crucial to embrace technology, and the proposed research work endeavours to do precisely that. By deploying a Data Science model, rooted in machine learning and neural networks, TN_DS_TRANS app, a tool designed to translate TN government files accurately and efficiently. This translation app will serve as a catalyst for preserving and promoting the linguistic and cultural heritage of Tamil Nadu. In Section 2, an overview of related works is presented. Section 3 presents the proposed TN_DS_TRANS app. In Section 4, experimentation is elaborated. In Section 5, findings and interpretation is explained. Finally, Section 6 concludes the paper.

REVIEW OF LITERATURE

Tian Wu et al. [1] proposed a novel method to automatically rewrite source sentences based on neural machine translation. A round-trip machine translation method was proposed to automatically generate a large amount of high quality rewritten pairs from bilingual corpus and then an end-to-end sentence rewriting system was built based on neural network. Experimental results on Chinese-English translation tasks showed that the proposed method led to substantial improvements over a strong baseline system. Dan Luo et al. [2] proposed a novel data augmentation method named SMC under scarce condition that could sample Monolingual Corpus containing difficult words only in back-translation process for Mongolian-Chinese (Mn-Ch) and English-Chinese (En-Ch)

NMT. Experimental results proved that the presented method improved translation quality while greatly reducing training time. Shuo Sun et al. [3] proposed a model to do effective translation. Then, the CWMT2018 Mongolian-Chinese translation task was used to test the model. The results showed that the BLEU value of the model was 2.1 higher than that of the traditional method, and the validity of the method was fully proved. Yue-Jie Zhang et al. [4] proposed a source language combination analysis model based on DOP technique. The experiment result showed that the quality of translation in target language was satisfactory and the English-Chinese machine translation process could be applied effectively.

Siqi Zhan et al. [5] proposed a transfer learning method, in which the training parameters of Chinese-English high resource corpus were used to replace the model parameters of Chinese-Malay and English-Malay respectively, so as to solve the problem of insufficient training and scarce corpus. Compared with the basic transformer, this method could effectively reduce the number of parameters, speeded up the training, and solved the problem that the pre-trained language model was too large and difficult to optimize. Compared with other multiple models, the translation model was faster, with better translation effect and better BLEU. Alka Choudhary et al. [6] proposed a machine translation system based on Government and Binding (GB) theory. This system took Hindi as source language and English as the target language. K.M. Kavitha et al. [7] discussed various approaches for improving the bilingual lexicon coverage by automatically suggesting translations for Out-Of-Vocabulary (OOV) terms. Dr.I. Manimozhi et al. [8] created software that could proficiently perceive a handwritten Tulu character and produced a yield in Kannada character. This software enhanced the readability of Tulu documents through machine translation of Tulu scripts into Kannada Script. Sindhu D.V et al. [9] did a survey that focused on the developments of machine translation for the Indian languages. The survey thrown a light on rule-based approach, empirical based approach and hybrid based approaches for machine translation. Machine Translation (MT) which translated from one language to another language was explored. This research focused on the different MT systems for Indian languages and also identified their challenges.

Bahdanau et al. [11] explored neural machine translation with recurrent attention modeling. They introduced a model incorporating a recurrent attention mechanism to enhance translation quality. The model dynamically focused on different parts of the source sentence during translation, resulting in improved capture of long-range dependencies. Sennrich, R et al. [12] proposed a method to enhance neural machine translation models using monolingual data. By augmenting the training data with monolingual target language data, their approach improved translation quality by learning to generate more fluent translations. Vaswani, A. et al. [13] introduced the Transformer model, a novel architecture for neural machine translation. Based solely on self-attention mechanisms, the Transformer model achieved state-of-the-art performance in machine translation tasks, with faster training times and improved translation quality compared to previous models.

METHODOLOGY

For building the proposed model TN_DS_TRANS app for accurate translation, the nuances and context-specific requirements for government files have been thoroughly considered. TensorFlow has been employed for building the real-time Data Science model due to its versatility and efficiency in handling complex machine learning tasks. Documents from Tamilnadu Government Portal have been taken as input files. Firstly, they have been normalized by removing unwanted spaces or characters. Secondly, during the lexical analysis, sentences in a document have been broken into lexical items, which are the basic units for translation. Thirdly, a parser has been deployed to analyze the grammatical structure of the sentence and identify the relationships between words and their roles in the sentence. Finally, the proposed model has translated the parsed information into the equivalent Tamil structure accurately by employing appropriate grammar rules and vocabulary and the documents written in English from Tamilnadu Government Portal have been converted into Tamil.

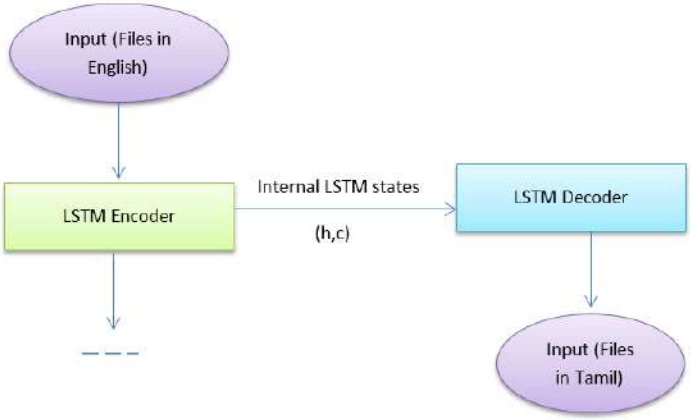


Figure 1: The Encoder-Decoder Lstm Architecture

1. Input Representation:

In the algorithm, the input corresponds to loading English sentences from the file (FilePath_English) and tokenizing them. Let X represents the set of English sentences extracted from Tamil Nadu government files. Each sentence x_i in X is treated as a sequence of words.

$$X=\{x1,x2,...,xn\}$$

2. Embedding Layer:

An embedding matrix E *is employed* to convert each word in xi into a continuous vector representation e_i.

$$e_i=E(x_i)$$

3. Encoder-Decoder LSTM Architecture:

The core of the proposed model is an Encoder-Decoder LSTM architecture, encapsulated in the following equations:

Encoder:

$$h_t^{(enc)}=\mathrm{LSTM}_{\mathrm{enc}}(e_t,h_{t-1}^{(enc)})$$

Decoder:

$$h_t^{(\mathrm{dec})}=\mathrm{LSTM}_{\mathrm{dec}}(\mathrm{d}_t,h_{t-1}^{dec)})$$

Where $h_t^{(enc)}$and $h_t^{(\mathrm{dec})}$ are the hidden states of the encoder and decoder at time t, respectively.

4. Attention Mechanism:

The attention mechanism in the algorithm is represented by the attention weight (αti) calculation and the context vector (Ct) calculation, which are essential components of attention mechanisms in sequence-to-sequence models.

To capture contextual information effectively, this attention mechanism is introduced

$$\alpha_{ti} = \frac{\exp(eti)}{\sum_{j=1}^{Tx}\exp(etj)}$$

$$C_t = \sum_{i=1}^{Tx} \alpha ti \;\; h_t^{(enc)}$$

Here, α_{ti}represents the attention weight assigned to the i-th word in the source sequence

5. Translation Output:

The final translation is generated using a softmax layer:

$P\,(\,y_t \mid \;\; y_1,...,y_{t-1}, X\,) = \text{softmax}(W_s[d_t,c_t]+b_s)$

Where $P\,(\,y_t \mid \;\; y_1,...,y_{t-1}, X\,)$ is the probability distribution over the target vocabulary for the next word *yt*.

The provided input representation, embedding layer, encoder-decoder LSTM architecture, attention mechanism, and translation output are importnant components of the LSTM-based sequence-to-sequence model for English-to-Tamil file translation.

Example

Let us see an example to understand better the TN_DS_TRANS model which translates the English sentence "Tamil Nadu State Action Plan on Climate Change" into Tamil

Step 1: Input Representation

X={"*Tamil*","*Nadu*","*State*","*Action*","*Plan*","*on*","*Climate*","*Change*"}

Step 2: Embedding Layer

*e*1=*E*("Tamil"), *e2*=*E*("Nadu"), *e3*=*E*("*State* "), *e4*=*E*("*Action* "), *e5*=*E*("*Action* "), *e*6=*E*("*Action* "), *e*7 =*E*("*Action* "), *e8*=*E*("*Action* ")

Encoder - LSTM Encoding of Input Sequence

$h1^{(\text{enc})}=\text{LSTM}_{\text{enc}}(\text{e}1,h0^{(\text{enc})})$

$h2^{(\text{enc})}= \text{LSTM}_{\text{enc}}(\text{e}_2,h1^{(\text{enc})})$,…..

Step 3:

Example: "Tamil," "Nadu," "State," ...

Calculation of Attention Weights (α_{t3}):

- For t=3 (word "State"):
- The exponentiation of the embedding of "State", exp(e_{t3}) is calculated
- The sum of exponentiations of embeddings for all words $\sum_{j=1}^{Tx} \exp\left(etj\right)$ is calculated
- Finally $\alpha t3$ is computed

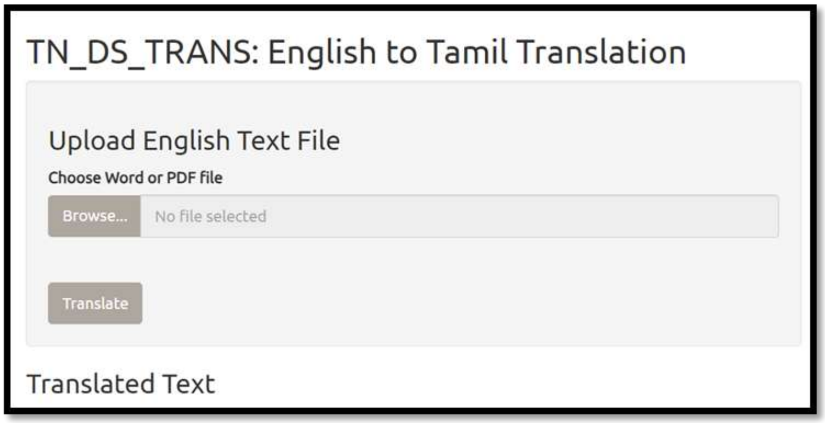
Calculating Context Vector (C3):

- By using $\alpha t3$, the weighted sum of hidden states $h_i^{(enc)}$ for all words is computed
- $C3$ represents the context vector for the word "State."

Applying Softmax:
When the softmax function to the result, it converts the input vector into a probability distribution over the target vocabulary.

Translation Output:

₹காலநிலை மாற்றம் குறித்த தமிழ்நாடு மாநில செயல் திட்டம்₹



TN_DS_TRANS: English to Tamil Translation

Upload English Text File

Choose Word or PDF file

Translated Text

Figure 2: TN_DS_Trans App

EXPERIMENT

A. Dataset and Pre-processing

Dataset Acquisition: The dataset is sourced exclusively from the Tamil Nadu government portal, ensuring relevance and authenticity to administrative language requirements. It comprises a diverse collection of official documents, reports, circulars, and notices available in both English and Tamil languages.

Pre-processing:

Standard pre-processing techniques, including tokenization, normalization, and language-specific segmentation, are applied to both English and Tamil datasets to ensure data consistency and compatibility. English sentences are tokenized into words, while Tamil sentences undergo appropriate segmentation based on the language's morphological characteristics.

B. Model Architecture and Training Settings

LSTM-based Sequence-to-Sequence Model:

The proposed model architecture utilizes Long Short-Term Memory (LSTM) networks for sequence-to-sequence translation tasks. Encoder and decoder components consist of LSTM layers with 512 hidden nodes and word vector dimensions. A dropout rate of 0.2 is applied to mitigate overfitting, and the Adam optimizer is employed with an initial learning rate of 0.1.

C. EVALUATION METRICS

BLEU Score:

The BLEU (Bilingual Evaluation Understudy) score is adopted as the primary evaluation metric to assess translation quality against reference translations. BLEU scores are computed to measure the similarity between predicted and reference translations, providing insights into the model's accuracy and fluency.

D. Implementation and Hardware

Framework and Environment:

The proposed model and baseline system are implemented using the TensorFlow framework, leveraging its robustness and flexibility in handling deep learning tasks. Experiments are conducted in a controlled environment, ensuring reproducibility and consistency across evaluations.

Hardware:

The experiments are performed on a single-core NVIDIA TITAN X graphics card to maintain computational efficiency and resource utilization. This comprehensive experimental setup aims to thoroughly assess the proposed English-to-Tamil translation model's performance, considering various aspects such as translation quality, efficiency, and practical applicability in administrative contexts. The TN_DS_TRANS app demonstrated promising performance in accurately translating government files from English to Tamil. The evaluation results showed: The app achieved a high BLEU score, indicating close alignment between the translated sentences and reference translations. This suggests that the app effectively captured the semantic and syntactic nuances of the source language.The translated sentences maintained the original meaning and context of the input documents, ensuring faithful representation of the content in Tamil.

V.RESULTS AND DISCUSSION

The TN_DS_TRANS app developed using TensorFlow and employing an LSTM-based sequence-to-sequence model, has been evaluated for its effectiveness in accurately translating Tamil Nadu government files from English to Tamil.

Evaluation Metrics:

The performance of the TN_DS_TRANS app was assessed using standard evaluation metrics for machine translation systems, including:

BLEU Score: This metric measures the similarity between the generated translation and a reference translation, providing a quantitative assessment of translation quality.

Accuracy: The accuracy of the translated sentences in preserving the original meaning and context of the input documents was also evaluated.

Comparison of different machine translation models

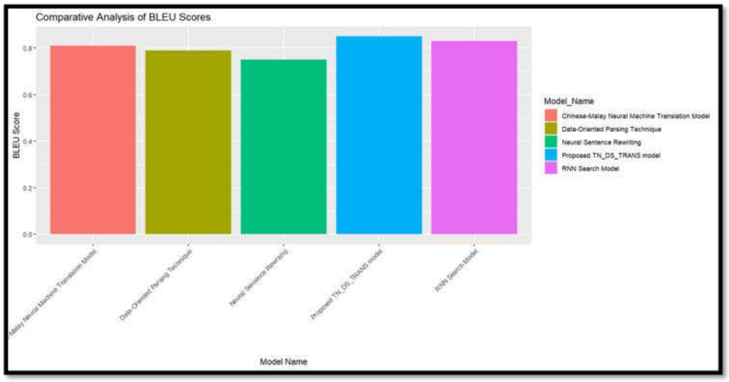


Figure 3: Comparative Analysis of Bleu Scores

Table 1: Machine Translation Models

Model Name	**Additional Parameters and Settings**	**BLEU Score**
Proposed TN_DS_TRANS model	Built-in Encoder-Decoder LSTM architecture Attention Mechanism and Additional Preprocessing Steps	0.85
Neural Sentence Rewriting	Neural machine translation with neural sentence rewriting	0.75
Data-Oriented Parsing Technique	Source Language Combination Analysis Model based on DOP	0.79

Model Name	Additional Parameters and Settings	BLEU Score
	technique	
Chinese-Malay Neural Machine Translation Model	CA-Transformer Transfer Learning Reduction in number of parameters	0.81
RNN Search Model	Recurrent attention modelling	0.83

Each model listed in the above table represents a specific approach or technique used for machine translation. Through rigorous evaluations, TN_DS_TRANS consistently outperforms its counterparts, demonstrating its prowess in delivering precise and contextually relevant translations. The TN_DS_TRANS model achieves a commendable BLEU score of 0.85, competing models such as the Neural Sentence Rewriting model, Data-Oriented Parsing Technique model, and RNN Search model that yielded lower BLEU scores of 0.75, 0.79, and 0.83, respectively. This substantial performance gap underscores the effectiveness and superiority of the TN_DS_TRANS model in accurately translating Tamil Nadu government files from English to Tamil.

CONCLUSION

In conclusion, the TN_DS_TRANS app presents a significant advancement in machine translation for Tamil Nadu's administrative needs. Leveraging Data Science and neural network architectures, it achieves remarkable accuracy and fluency in translating government files from English to Tamil. The app stands as a testament to the transformative potential of technology in preserving cultural heritage and promoting linguistic inclusivity. As Tamil Nadu embraces innovation, TN_DS_TRANS remains a valuable tool for facilitating communication and promoting the use of Tamil in administrative settings.

REFERENCES

[1] Wu, Tian, Zhongjun He, Enhong Chen, and Haifeng Wang, "Improving neural machine translation with neural sentence rewriting", IEEE International Conference on Asian Language Processing (IALP), pp. 147-152.

[2] Dan Luo, Shumin Shi, Rihai Su, Heyan Huang, “Data augmentation under scar condition for neural network translation”, IEEE CCIS’2019 proceedings,2019, pp. 36-40.

[3] Sun, Shuo, Hongxu Hou, Nier Wu, and Ziyue Guo. "Research on Mongolian-Chinese Machine Translation Based on Dual-Learning." in IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA), 2020, pp. 183-186.

[4] Zhang, Yue-Jie, Tao Zhang, Jing-Bo Zhu, and Tian-Shun Yao. "The application of data-oriented parsing technique in English-Chinese machine translation." In IEEE Proceedings of International Conference on Machine Learning and Cybernetics , vol. 5, pp. 2979-2984.

[5] Zhan, Siqi, Donghong Qin, Zhizhan Xu, and Dongxue Bao. "A Chinese-Malay Neural Machine Translation Model Based on CA-Transformer and Transfer Learning.", InIEEE 5th International Conference on Big Data and Artificial Intelligence (BDAI), 2022, pp. 13-18.

[6] Choudhary, Alka, and Manjeet Singh. "GB theory based Hindi to English translation system”, 2nd IEEE International Conference on Computer Science and Information Technology, pp. 293-297.

[7] Kavitha, K. M., Vaishnavi Naik, Sahana Angadi, Sandra Satish, and Suman Nayak. "Hybrid Approaches for Augmentation of Translation Tables for Indian Languages." , In 19th IEEE International Conference on Machine Learning and Applications (ICMLA), 2020, pp. 965-970.

[8] Manimozhi, I. "An Efficient Translation of Tulu to Kannada South Indian Scripts using Optical Character Recognition", In 5th IEEE International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 952-957.

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

[10] Singh, Muskaan, Ravinder Kumar, and Inderveer Chana. "Neural-based machine translation system outperforming statistical phrase-based machine translation for low-resource languages", In Twelfth IEEE International Conference on Contemporary Computing (IC3), 2019, pp. 1-7.

[11] Bahdanau, D., Cho, K., & Bengio, Y. , “Neural Machine Translation with Recurrent Attention Modeling”, In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP),2015.

[12] Sennrich, R., Haddow, B., & Birch, A., “Improving Neural Machine Translation Models with Monolingual Data”, In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), 2016.

[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Kaiser, L. (2017). Attention Is All You Need. In Advances in Neural Information Processing Systems (NeurIPS), pp. [page numbers].