Efficient use of machine learning models to evaluate the parametric performance of the DL models for language translation from Telugu to Hindi.

Shail garg
Department of Computer Science and Engineering
Amrita School of Computing, Amrita Vishwa Vidyapeetham,
Bengaluru, India.
bl.en.u4cse22254@bl.students.amrita.edu

Yerukola Gayatri
Department of Computer Science and Engineering
Amrita School of Computing, Amrita Vishwa Vidyapeetham,
Bengaluru, India.
bl.en.u4cse22267@bl.students.amrita.edu

A. Surya Kausthub
Department of Computer Science and Engineering
Amrita School of Computing, Amrita Vishwa Vidyapeetham,
Bengaluru, India.
bl.en.u4cse22287@bl.students.amrita.edu

***Abstract*—Due to the language differences involved, there are lots of issues while translating Telugu which is a low resource language to Hindi, which is one of the most widely used languages in India. This research presents the different models of advanced machine translation such as OpenNMT, Fairseq machine translation, Helsinki-NLP/OPUS-MT, and BART or Bidirectional and Auto-Regressive transformers to solve these challenges. It is for this reason that the study expects to engage in systematic experimentation of these models under various parameter conditions with the overall aim of ascertaining their parametric ability to produce contextually correct translations. The primary aim here is to determine which models and approaches lead to improvement of the quality of translations, thus helping improve on the existing solutions for underrepresented languages such as Telugu. Dubbed as The Roadmap for enhancing accessibility and communicational opportunities in low-resource languages with MT, the findings of this research will positively impact the particular field.**

***Index Terms*—Telugu to Hindi Translation,Low-Resource Language,OpenNMT,Fairseq,Helsinki-NLP/OPUS-MT,BART**

I. Introduction

Language translation is fundamental in NLP since it helps in language barriers hence assists in enhancing global relations in terms of culture, economy, language among other relations. The process of translating from Telugu to Hindi, or for any other pair of languages that are phonetically, syntactically and, to an extent semantically dissimilar entails much more than mapping twenty-six letters onto homophones, fingers onto typewriter keys, or concepts onto a correlate array of symbols. Though both the Telugu, a Dravidian language of Andhra Pradesh in India, and Hindi – one of the most extensive languages in India – share no similitude in their script, grammar, syntax, semiotic systems, translating from one to the other is a challenge.

The first difficulty which appears when translating from Telugu to Hindi is the absence of large parallel corpora required to train deep learning models. Telugu, which is a low resource language having limited availability of digital text data is a major hurdle as most of today's MT systems require large amount of parallel data for achieving high level results. Even fewer involve real subject matter, let alone academic, research, legal or medical subject matter, making the translation task even more challenging especially when facing low resources settings.

It is interesting to note that in recent years deep learning has imposed more flexible approaches to address the imbalance between the languages. Players as OpenNMT, Fairseq, Helsinki-NLP/OPUS-MT, and BART (Bidirectional and Auto-Regressive Transformers) have emerged in this area, with each of them carrying its peculiarities into this area. Fairseq among them is famous for its solid sequence-to-sequence structures and flexibility of the framework, which makes both models fit for fine-tuning on particular translation tasks. Helsinki-NLP/OPUS-MT, for instance, uses a range of multilingual models which means that low resource languages, such as Telugu to Hindi translations can also be made. Since BART is both bidirectional and autoregressive it presents strong capabilities for translations, especially for more complicated language pairs due to its fluency and contextual relevance.

However, like any other research, each model has its pros and cons when it comes to low-resource languages especially the Telugu language. It is therefore important to know how these models behave depending on different conditions of use, the different parameters and adjustments that one can make in an aim to produce the best translations. This study is carried on by analyzing the architectural details, tuning techniques, and computational complexity integral to these models to improve the Telugu-to-Hindi translation. The knowledge obtained will be helpful to create more effective machine translation approaches for low-resource languages, hence providing paths toward enhanced interaction in different settings.

II. Literature survey

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, Fu Lee Wang [1] Did a Critical Review and Assessment While preserving performance, parameter-efficient fine-tuning (PEFT) techniques use less memory and fewer fine-tuning parameters. They are utilized in cross-lingual transfer, backdoor attack protection, and multi-task learning. For best results, they incorporate new parameters or combine different PEFT methodologies. Future studies will examine PEFT approaches in multi-modal learning and computer vision, as well as improve their efficiency and explainability.

Yasir Abdelgadir Mohamed; Akbar Khanan; Mohamed Bashir; Abdul Hakim H. M. Mohamed; Mousab A. E. Adiel [2] Worked on machine translation developments, with a special emphasis on neural machine translation (NMT) systems. It emphasizes how deep learning and artificial neural networks can be used to increase the accuracy, efficiency, and quality of translation. The study highlights the need for additional research to improve translation quality and address cultural nuances, as well as the significance of assessing machine translation systems using both automated metrics and human evaluations.

Manuel Eugenio Morocho Cayamcela,Wansu Lim [3] highlights advances in neural machine translation and deep learning, as well as ongoing conversations about inclusivity and cultural sensitivity, in order to examine how AI is changing language translation. Fuzzy algorithms tackle semantic ambiguity and linguistic variability, while feature extraction, intelligent recognition, and maximum-entropy principles improve language context awareness. AI is transforming translation. .

Kexin Zhang [4] Explains how back-translation and cross-lingual embeddings, along with creative strategies, enhance translation quality. It highlights the significance of resolving issues with NMT training techniques for more effective and precise translation systems by showing notable gains over conventional unsupervised systems. With assessments concentrating on evaluating translation quality through automated and human-based methodologies, the study also emphasizes machine translation as a useful tool in increasing efficiency and lessening the workload on translators.

Teddy Mantoro; Jelita Asian; Media A. Ayu [5] Applied sequence IRSTLM translation parameters and pruning to enhance a statistical machine translator's translation performance. It talks about how difficult it is to translate and offers a method that can still create accurate translations without requiring a deep understanding of the target language. The study highlights the significance of user interface, customisation, and pruning in machine translation while evaluating 28 distinct IRSTLM language modeling factors. The suggested strategy outperforms conventional approaches that rely on linguistic expertise, and the results reveal promising outcomes.

Kahler, B., Bacher, B. and Jones, K.C. [6] proposed a technique which addresses character conversion problems and enhances machine translation (MT) accuracy by employing ISO character mapping to improve reliability. It minimizes mistranslations and absurd results by drastically lowering errors when converting characters from languages like Arabic, Asian, and Cyrillic to Western scripts. Users consequently gain a deeper comprehension of foreign-language online information. The authors emphasize how easily open-source MT tools may be integrated with their suggested developer solution. This advancement in translating technology has wider ramifications for improving cross-cultural relationships and lowering obstacles to communication. The study recommends more investigation to hone these techniques and look at other ways to enhance MT systems.

Sun, S., Hou, H.X., Yang, Z.H. and Wang, Y.S[7] created an innovative method to improve translation for languages that are less well-known and have little data. To increase translation accuracy, especially when a substantial quantity of bilingual data is absent, it leverages the powerful pre-trained model CeMAT. How to avoid the model from repeatedly making the same mistakes is one of the primary problems that is mentioned. They tackle this by presenting an approach that lets the model grow from its errors. Additionally, they offer a clever training strategy that modifies based on the data and the confidence level of the model, particularly helpful for languages with limited resources. Their experiments show that these approaches translate substantially better, proving the effectiveness of pre-training in conjunction with this innovative learning strategy for low-resource languages.

III. Methodology

In this section, we describe the application of regression and clustering techniques to analyze the relationship between Telugu and Hindi texts.

A. Regression Analysis

1) Data Collection and Preprocessing: We collected parallel pairs of Telugu-Hindi texts. The dataset was cleaned by removing rows with missing values. The Telugu text was used as the independent variable X, and the Hindi text length was the dependent variable y. To handle textual data, we applied TF-IDF vectorization to transform the Telugu text into numerical features.

TF-IDF assigns weights to words based on their frequency and uniqueness across the dataset. This captures important linguistic features while reducing noise from commonly occurring words. We split the dataset into an 80-20 training and testing set using `train_test_split`. The training set was used to train the regression model, and the test set was reserved for evaluating its performance.

2) Model Training: We employed a linear regression model to predict the length of the Hindi translation based on the TF-IDF vectors of the Telugu text. The model was trained using the training set (X_{train}, y_{train}), where X_{train} consisted of TF-IDF-transformed Telugu text and y_{train} was the corresponding Hindi text length.

3) Model Evaluation: We evaluated the model using the following metrics:

- Mean Squared Error (MSE): Measures the average squared differences between predicted and actual values.
- Root Mean Squared Error (RMSE): The square root of MSE, providing an error metric in the same units as the predicted values.
- Mean Absolute Percentage Error (MAPE): Measures the average error in percentage form.
- R^2 Score: The proportion of variance in the dependent variable that is predictable from the independent variables.

B. K-Means Clustering

1) K-Means Clustering: Clustering was applied to group Telugu texts based on similarities. We used the K-Means algorithm, which assigns data points to k clusters based on the closeness of each point to the cluster center. Initially, we set $k = 2$.

The Telugu text was transformed into numerical features using TF-IDF vectorization, and the K-Means algorithm clustered the data based on these features.

2) Clustering Evaluation: The quality of the clusters was evaluated using the following metrics:

- Silhouette Score: Measures how similar a data point is to its own cluster compared to other clusters.
- Calinski-Harabasz Index: Calculates the ratio of between-cluster dispersion to within-cluster dispersion.
- Davies-Bouldin Index: Measures the average similarity between clusters, where lower values indicate better clustering.

3) Determining Optimal Clusters: To determine the optimal number of clusters, we used the Elbow Method, plotting the distortion (sum of squared distances between each point and its cluster center) for various values of k. The ideal k corresponds to the point where the distortion decrease becomes negligible.

IV. Results and Discussion

A. Results of Regression Model

1) Training Performance: The linear regression model achieved the following performance on the training set: These results indicate that the model fit the training data well, with a high R^2 score and low MSE and RMSE values.

2) Test Performance: However, the performance on the test set was significantly worse:

- MSE: 460,592.56
- RMSE: 678.67
- MAPE: 5,568%
- R^2 Score: 0.41

The drastic drop in performance on the test set indicates severe overfitting, where the model memorized the training data but failed to generalize to unseen data.

3) Discussion: The results highlight the limitations of using linear regression for predicting text translation length. The oversimplified assumption of a linear relationship between Telugu text and Hindi translation length could not capture the complexity of language translation. More advanced models, such as Support Vector Machines (SVM), Random Forest, or Neural Networks, may offer better performance.

TABLE I
Regression Model Performance

Metric	**Training Set**	**Test Set**
MSE	12.35	460,592.56
RMSE	3.51	678.67
MAPE	4.12%	5,568%
R^2 Score	0.99998	0.41

B. Clustering Analysis Results

1) Initial Clustering ($k = 2$): For $k = 2$, the clustering results were as follows:

- Silhouette Score: 0.0044
- Calinski-Harabasz Index: 1.56
- Davies-Bouldin Index: 1.67

These metrics suggest poorly defined clusters, with most data points being equidistant to multiple clusters.

2) Optimal Number of Clusters: Using the Elbow Method, we determined that the optimal number of clusters is $k = 3$. However, the clustering performance remained subpar, likely due to the limitations of TF-IDF vectorization in capturing deeper semantic relationships.

3) Discussion: The poor performance of K-Means clustering suggests that TF-IDF is not well-suited for grouping texts based on semantic content. Word embeddings such as Word2Vec, GloVe, or BERT could capture richer linguistic features and improve clustering performance. Additionally, hierarchical clustering or DBSCAN might offer better results.

TABLE II
K-Means Clustering Performance

Metric	**k=2**	**k=3**
Silhouette Score	0.0044	0.0123
Calinski-Harabasz Index	1.56	2.74
Davies-Bouldin Index	1.67	1.45

C. General Observations

Key insights from the experiments include:

- Linear regression fails to capture the complexity of text translation tasks when using word count as the target variable.
- K-Means clustering with TF-IDF vectorization performs poorly at segmenting similar texts. More sophisticated feature extraction techniques, such as word embeddings, could yield better results.

V. Conclusion

In this study, we applied regression and clustering techniques to Telugu-Hindi parallel text data. While the linear regression model fit the training data well, it failed to generalize to new data, indicating overfitting. Similarly, K-Means clustering with TF-IDF vectorization produced poorly defined clusters, suggesting that TF-IDF does not capture all the necessary linguistic features. Future work should explore more advanced models, such as deep learning and word embeddings, to better account for the complexities of language translation and text clustering.

References

[1] Xu, L., Xie, H., Qin, S.Z.J., Tao, X. and Wang, F.L., 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. arXiv preprint arXiv:2312.12148..

[2] Mohamed, Y.A., Khanan, A., Bashir, M., Mohamed, A.H.H., Adiel, M.A. and Elsadig, M.A., 2024. The impact of artificial intelligence on language translation: a review. Ieee Access, 12, pp.25553-25579.

[3] Cayamcela, M.E.M. and Lim, W., 2019, February. Fine-tuning a pre-trained convolutional neural network model to translate American sign language in real-time. In 2019 International Conference on Computing, Networking and Communications (ICNC) (pp. 100-104). IEEE.

[4] Zhang, K., 2021, May. Application of Pretrained Models for Machine Translation. In 2021 International Conference on Communications, Information System and Computer Engineering (CISCE) (pp. 849-853). IEEE.

[5] Mantoro, T., Asian, J. and Ayu, M.A., 2016, October. Improving the performance of translation process in a statistical machine translator using sequence IRSTLM translation parameters and pruning. In 2016 International Conference on Informatics and Computing (ICIC) (pp. 314-318). IEEE.

[6] Kahler, B., Bacher, B. and Jones, K.C., 2012, July. Language translation of web-based content. In 2012 IEEE National Aerospace and Electronics Conference (NAECON) (pp. 40-45). IEEE.

[7] Sun, S., Hou, H.X., Yang, Z.H. and Wang, Y.S., 2023, June. Multilingual Pre-training Model-Assisted Contrastive Learning Neural Machine Translation. In 2023 International Joint Conference on Neural Networks (IJCNN) (pp. 01-07). IEEE.