Advancing Roman Urdu to Urdu Transliteration using Machine Learning Techniques

By Ahsan Ahmad & Mohsin Ali Ahmad***

**Senior Machine Learning Engineer*

***Student*

ABSTRACT

Roman Urdu is widely used by people for communication on social media platforms and daily messaging especially in Pakistan due to which writing Urdu is difficult for them. In this paper, we research different models and compare their results on a dataset of approximately 6.5 million sentences. Lastly, we suggest different modifications in the architecture of the Transformer model (which give us the best results) to improve the BLEU score of the Roman Urdu to Urdu transliteration to increase the generalization and accuracy of the transliteration enabling the transliteration of the sentence according to its context.

INDEX TERMS

Roman Urdu [1], [2], Romanization [3], Urdu script [4], [5], Transcription [6], [7], Phonetic conversion [8], [9], Roman Urdu to Urdu mapping [10], [11], Script conversion [12], [13], Romanized Urdu [14]

1. INTRODUCTION

Machine translation, sometimes referred to by the abbreviation MT is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one language to another. The predefined machine translation techniques can also be used for transliteration purposes which means transferring a word from the alphabet of one language to another. Nowadays, Roman Urdu is widely used as a source of communication in Pakistan.

People have difficulty reading and typing Urdu scripts. We can observe that in our daily tasks i.e.; in writing messages, informal letters and emails, posting something on social media, etc, Roman Urdu is widely used instead of Urdu script. Moreover, we are used to typing English alphabets from the keyboards and face difficulty in finding the Urdu alphabets from the keyboard. Therefore, we will apply different techniques for improving the quality of transliteration to assist people in writing the Urdu script without errors. After improving the quality of the transliteration, we would develop an Android Application for this purpose. It would assist people in learning and writing Urdu script easily as the application would be with them wherever they go on their phones. The application would have a simple UI and would be easy to use for people of any age.

Some web applications are currently present for the transliteration purpose but they are either in the learning mode or are not able to give accurate transliteration according to the context of the sentence. Therefore, for learning purposes, the application must give accurate transliteration results. We would improve the quality of Roman Urdu to Urdu transliteration by increasing the quality and quantity of the current data set and by using modern deep learning techniques enabling us to produce better and more generalized results. Some of the challenges that we would face while building the model for transliteration purposes are:

- Transliteration of rare words
- Transliteration of unseen words
- Transliteration of the same word having different spellings in Roman Urdu Transliteration according to the context

We will be applying deep learning techniques to overcome these challenges and will be documenting our findings. We will also be developing an Android application that would transliterate Roman Urdu sentences into Urdu script sentences. The application would assist people in typing Urdu script. It would save them time and effort. Moreover, it would be easier for them to learn Urdu script by using this application. Therefore, it is necessary for this application to gives transliteration quality nearer to human language. In this project, we perform a comparison of different models used for machine translation RNN+LSTM, seq2seq, and Transformer Model and compare their performances over our data set. We perform various techniques to increase the amount and diversity of our data set to generalize the transliteration results.

2. OBJECTIVES

Our system has the following objectives:

- To enhance the quality and expand the quantity of the dataset for Roman Urdu to Urdu transliteration.
- To conduct a comparative analysis among three machine translation models RNN+LSTM [15], [16], seq2seq [17], and Transformer Model [18], [19], to evaluate their performance.
- To implement various architectural modifications to optimize the transliteration process and achieve superior results.

3. LITERATURE SERVEY

Deep neural networks are powerful for solving complex tasks, but they can overfit when training on smaller data sets or longer periods. To address this issue, a method called

"Dropping Out" is proposed, which involves dropping randomly selected neurons in the network. This technique works better than other techniques, such as max-out, max-pooling, and dropping-outs. Dropouts help avoid overfitting and make models more general, but they introduce noise in the gradient. To avoid this, max-norm regularization is used instead. [20]

BLEU (Bilingual Evaluation Understudy) is a metric used by text miners to assess the quality of machine translations. It uses a weighted average of variable length phrase matches against reference translations, creating n-gram models to check their presence in the candidate translations. The BLEU score ranges from 0-1, with translations only reaching a BLEU score of 1 when they are identical to the reference translation. This is rare in real life, and a human translator scored a BLEU score of 0.3468 against 4 reference sentences and about 500 sentences in the corpus. BLEU is quick, inexpensive, easy to understand, and language independent, correlated with human evaluation, and has been widely adopted. However, MT systems can over generate reasonable words, resulting in improbable but high-precision translations. [21]

This paper presents an extended encoder-decoder system for English-to-French translation, which encodes the input sentence into multiple encoded vectors, producing better results even on longer sentences. The system uses hidden states produced by one hidden layer in the prediction of the next target word and in the prediction of the next hidden layers. This allows the system to remember the previously produced targeted word and the context of the previous hidden states. The decoder calculates the conditional probabilities of translations using a non-linear, multilayered function. The encoding part uses bidirectional RNN models, with forward RNNs reading an ordered input sequence and backward RNNs reading a reversed sequence. The decoding part uses the previously calculated hidden states and predicted targeted values to produce a translation against the input word x. The proposed model mainly provides a solution to translating long sentences, creating multiple vectors from the input sentence, and creating a sub-vector for further processing. [22]

This research paper explores machine translation from English to French using neural networks. The method involves using an LSTM to extract a large context vector and then output the translation using a recurrent neural network language model. The WMT-14 English to French dataset was used for training and testing. The model achieved a BLEU score of 34.81 with a vocabulary of 80k words. The authors used deep LSTMs with a limited vocabulary to outperform shallow LSTMs. They penalized translation for out-of-vocabulary words and achieved a better BLEU score than the state of the art. However, they could have addressed the out-of-vocabulary word problem by increasing the target language's vocabulary or creating a dictionary for future translations. [23]

This paper proposes a technique to address the rare word problem in machine translation by training an NMT system on data augmented by the output of a word alignment algorithm. The system emits the position of an out of vocabulary (OOV) word in the target sentence, which helps in translating each word using a dictionary as the post-processing step. Three models were implemented: the Copyable Model, the Positional All Model, and the Positional Unknown Model. The model was trained on the French vocabulary of the 40K or 80K most frequent French words on the target side and the 200K most frequent English words on the source side. The model improved BLEU scores by 2.3-2.8 BLEU points and 1.6-1.9 BLEU points, with some penalties due to incorrect alignments and wrong entries in the dictionary. [24]

The authors developed a machine translation using a sequence-to-sequence model, consisting of an encoder and decoder. The model accommodates the heterogeneity of source and target languages, regardless of their orientation. LSTM cells were used to overcome gradient vanishing problems and handle long sentences. The model is heterogeneous, capable of training on any language type and handling sentences up to 10 words. However, it degrades on sentences longer than 10 words and cannot correctly translate bi-words like Islamabad. The model's accuracy gap needs improvement. [25]

Bag-Of-Words (BOW) is a common vector representation method for text representation, but it struggles to keep the context of words. Bag-of-n-grams consider word order in a short context but suffer from data sparsity and high dimensionality, making it computationally expensive. To overcome this, a new technique called "Paragraph Vectors" has been proposed. The authors prepare two vectors: word vectors of the whole corpora using techniques mostly BOWs and a paragraph vector of a specific paragraph. These vectors are prepared using unsupervised learning and are trained using stochastic gradients using backpropagation. The paragraph vectors are trained using stochastic gradients and averaged or combined to form a larger vector. They address some of the weaknesses of BOWs models, such as preserving semantics and considering word order in small contexts. However, the model is computationally expensive due to the computation of many weights. The proposed method is better than bag-of-n-grams models, as it preserves more information about the paragraph, including word order. [26]

The author web-crawled the proceedings of the European Parliament and trained a model for 110 language pairs for statistical machine translation. The corpus consisted of 30 million words from the official languages of the European Union, including Danish, German, Greek, English, Spanish, Finnish, French, Italian, Dutch, Portuguese, and Swedish. The process of translation for other languages began with the inclusion of other countries in the EU. The authors pre-processed the corpora by obtaining raw data, extracting parallel chunks, normalizing and tokenizing the data, and mapping each sentence to its translation in other languages. Challenges faced in data normalizing and tokenization included identifying abbreviations and possessive markers and simplifying computations. The BLEU scores were calculated, showing that languages with higher morphological differences are easier to translate. Clustering techniques can improve translation quality for languages with higher morphological differences. [27]

4. METHODOLOGY

A. DATASET DESCRIPTION

1) Data Collection

The data set was collected and expanded using techniques such as data scraping from novels and other Urdu websites [28], [29], crowd-sourcing from students in a lab, and data generation [30], [31] by including different spellings for the same word in the Urdu script, such as Roman Urdu sentences against "raha."

2) Data Diversity

Data diversity was added through techniques like adding different spellings of Roman Urdu for the same word and adding tweets and formal sentences to cover difficult and daily usage words.

3) Data Cleaning

Data from various sources was cleaned manually and coded to remove slang words, script words, and garbage symbols to ensure accuracy and maintain clean sentences.

4) Transliteration of Collected Data

The cleaned data was transliterated into Roman Urdu script using APIs [32], [33], which were then reviewed and manually corrected.

B. DETAILS OF MODEL

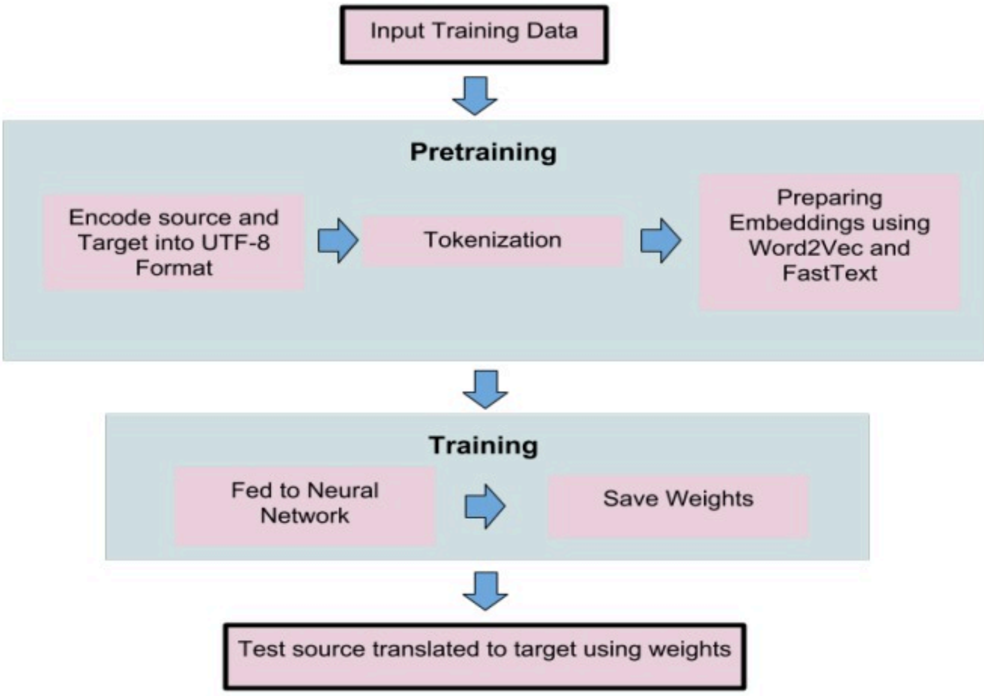
1) Model Working

The model, Roman2Urdu, was trained using specific parameters and ratios. The model was registered in tensor2tensor [34] using the standard transformer model registration method. The hyperparameter was set to 20000, which was initially set to 4000 for German to English translation. A 90/10 split was used for training and evaluation, reducing overfitting. The model's architecture used attention-based layers with 4 hidden layers, 128 size, and 512 filter size. Drop-out layers were introduced to prevent overfitting and a learning rate of 0.01. Parallel corpora were used for training.

2) Training Steps

The Transformer model is trained on a data set of 6.5 million Roman Urdu sentences, which are transliterated into Urdu script. The data is converted to UTF-8 format, tokenized, and prepared for embedding. The Word2Vec algorithm extracts relevant embeddings. After pre-training, the model is fed into a neural network for training. The weights saved during training are used for transliteration at runtime. The BLEU score [35] is computed to estimate the accuracy of the transliteration.

The training process is also explained in the figure below:



3) Transliteration Steps

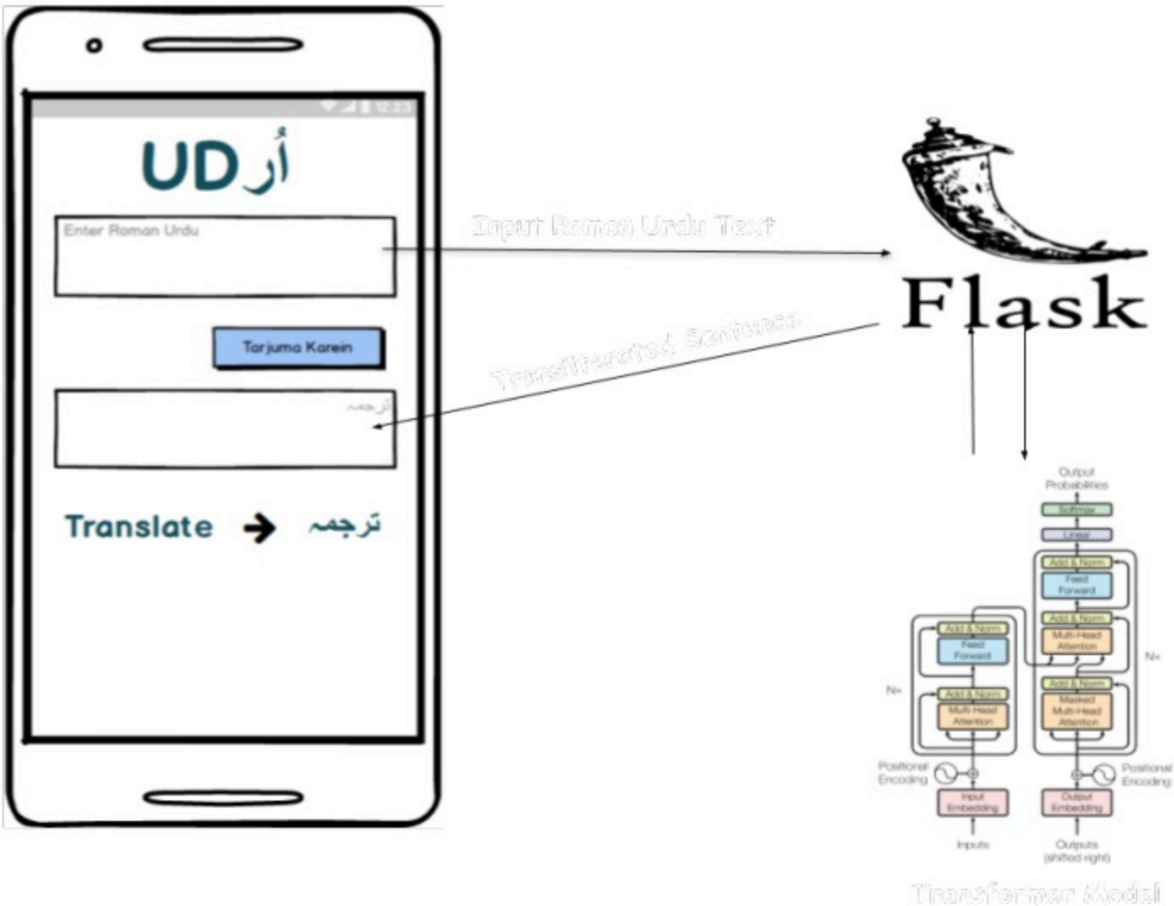
The transliteration steps are somewhat the same as the training steps of the data set.

- The Roman Urdu sentences to be transliterated are written in a file. The sentences are given to the model for transliteration.
- Each of the sentences is then converted into UTF-8 format[36], tokenization [37], [38] is performed to extract individual words and embedding is prepared.
- After that, the sentences are fed to the neural network [39]and transliterated using the saved weights.
- The transliteration with the highest probability is then decided for the sentence. Also if any of the word is not present in the data set, its transliteration is produced using one-to-one mapping of the Roman Urdu word to the Urdu script alphabet.

C. WORKING OF ANDROID APPLICATION

ISSN 2582 8088
Volume 5 Issue 2 – March April 2024

An Android app, written in Kotlin [40], [41], allows transliteration from Roman Urdu to Urdu. Connected to a Flask web server [42], the app simplifies loading the model into the same application. The diagram below gives a good idea about how the application works:



1) **Steps**

- User writes in the text box and clicks the translate button.
- The application then calls an API with the payload (Roman Urdu data).
- The flask server then receives data and sends it to the model.
- The model then transliterates the data using the saved weights and sends back the result to Flask.
- Flask then returns the transliterated result to the application.
- The Application then displays the result to the user on the screen.

ISSN 2582 8088
Volume 5 Issue 2 – March April 2024

5. IMPLEMENTATION

A. EXPERIMENTAL SETUP

1) Effect of Data Diversity

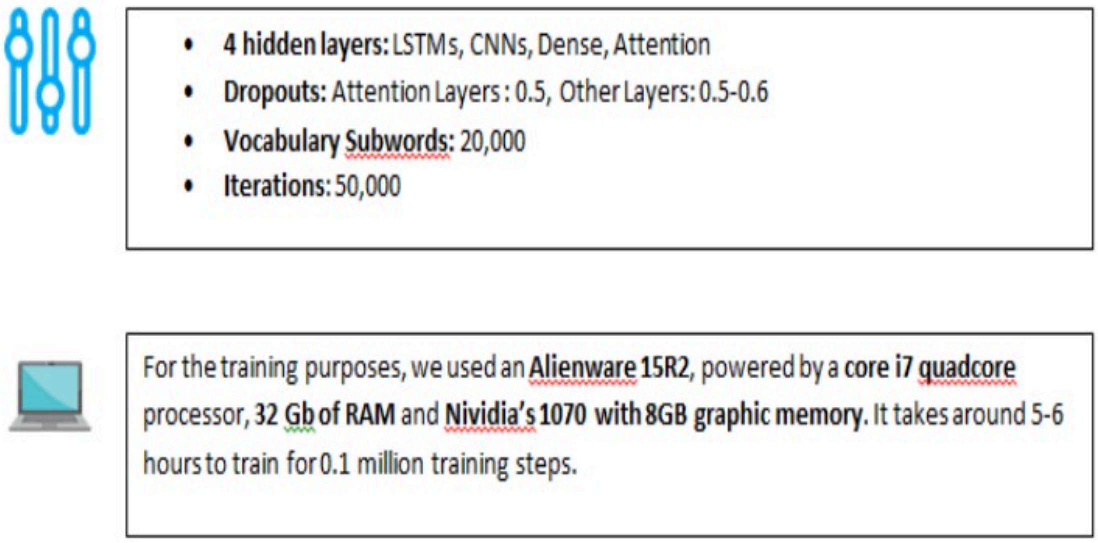
The BLEU score initially lacked contextual understanding in transliteration due to a large dataset of 1 million. To improve accuracy, the data set was expanded to include diverse sources, age groups, informal and formal data, and poetry. The data generation technique improved the model's understanding of different contexts.

2) Parameter Tuning

The transformer model was used for training due to its superior transliteration results. The model had 4 attention-based hidden layers, 128 filters, and 512 filters. To avoid overfitting, drop-out layers were introduced, dropping 50% of connections randomly. The learning rate was set at 0.01. The model had 20000 vocabulary sub-words for each source and target data set. A 90/10 split was used for training and cross-validation, with updated weights saved in a checkpoint file. The model trained on 6.5 million data sets from crowd-sourced resources.

3) System Details

The training process utilized an Alienware 15R2 with a core i7 quadcore processor, 32GB RAM, and 8GB graphic memory, taking 5-6 hours for 0.1 million steps. The final parameter and system details are shown in the diagram below:



B. EVALUATION

1) Evaluation Metric

The BLEU score is used to evaluate the quality of transliteration, focusing on the similarity between the model's translated text and the original transliterations. This metric is chosen for its accuracy and compatibility with previous machine translation research, making comparison easier.

2) Comparison of models

- **Seq2Seq**

  Initially, we studied the seq2seq model. It gave fairly good results in some of the cases. But for some cases like in the case of long sentences and rare words it didn't give any commendable results. Seq2Seq2 model gave a BLEU score of 48% when trained on around 1 million data points.

- **Tensor2Tensor**

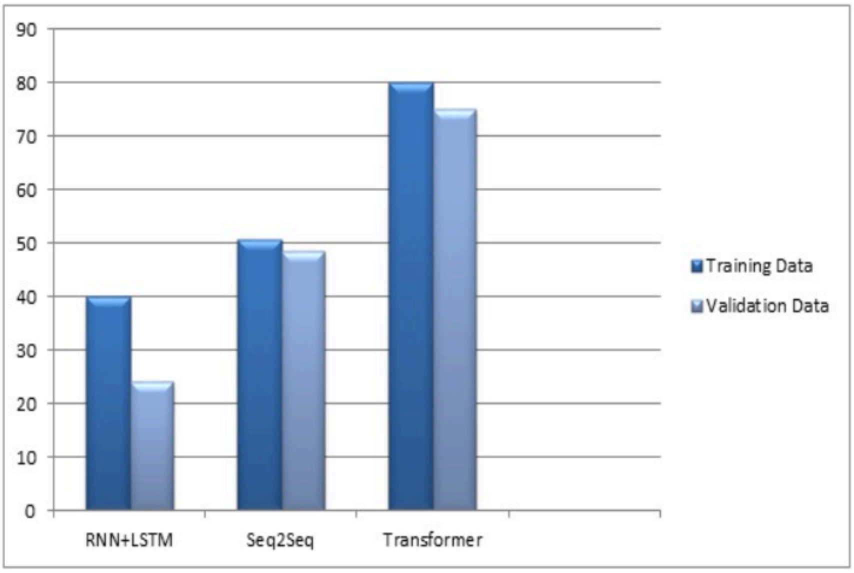
  After trying the Seq2Seq model, we studied another attention-based model built mostly for translation purposes. This countered the flaw of long sentences being faced in the Seq2Seq model and gave us good results irrelevant to the length of the sentence. The BLEU score achieved from the tensor2tensor model was surprisingly around 80%. Tensor2Tensor model tried to build a contextual understanding between the words which gave good results. Also, it tried to cater to the problem of ultra-rare and unknown words by trying to make an alphabetical relation.

- **Custom Model**

  We also tried to make our model using simple RNNs and LSTMs. However, it did not give any good results as the BLEU score was only around 20% also it did not build any sort of contextual understanding.

The comparison of the BLEU scores of the three models is shown below in the graph:

ISSN 2582 8088
Volume 5 Issue 2 – March April 2024



6. RESULTS AND DISCUSSION

A. COMPARISON OF RESULTS

The best results were achieved using the Transformer model. The comparison of the results of the two best models i.e. seq2seq and Transformer model is shown in the table below:

		BLEU Score	
Model	**Data**	**Training Data**	**Validation Data**
Seq2seq	1 Million	50.68	48.6
Tensor2tensor	6.5 Million	80	75

Some of the Roman Urdu sentences were given as input to both the models as well as a widely used website "ijunoon" to compare the transliteration results given by each of them. Some of the results worth noting are given below:

Roman Urdu Sentence	Transformer	Seq2Seq	ijunoon
iqbal musalmano ko phir isi akhuwat e islami ki taraf lotnay ki talqeen karte hain	اقبال مسلمانوں کو پھر اسی اخوت اسلامی کی طرف لوٹنے کی تلقیں کرتے ہیں	اقبال مسلمانوں کو پھر اسی اخوت اسلامی کی طرف لُوٹنے کی تلقین کرتے ہی	اقبال مسلمانوں کو پھر اسی اخوت اسلامی کی طرف لوتنے کی تلقین کرتے ہیں
Inhen bachpan hi se ilm o adab aur drama se dilchaspi thi	انہیں بچپن سے ہی علم و ادب اور ڈرامہ سے دلچسپی تھی	انہیں بچپن سے ہی علم ادب اور ڈرامہ سے دلچسپی تھی	انہیں بچپن ہی سے علم و ادب اور ڈرامہ سے دلچسپی تھی
imla mein koi tabdeeli aisi tajweez na ki jaye	املا میں کوئی تبدیلی ایسی تجویز نا کی جائے	املا میں کوئی تبدیلی ایسی تجویز نہ تھی جائے	املا میں کوئی تبدیلی ایسی تجویز نا کی جائے
Ahal islam ke nazdeek khuda se dua maangna ibadat mein shaamil hai	اہل اسلام کے نزدیک خدا سے دعا مانگنا عبادت میں شامل ہے	اہل اسلام کے نزدیک خدا سے دعا قدروں تلاش میں شامل ہے	اہل اسلام کے نزدیک خدا سے دعا مانگنا عبادت میں شامل ہے

Analysis of the Compared Results:

As we can see the transformer model most of the time gives accurate results. Also, whenever the transformer model does not give an accurate result, it is very near to the correct result in the 4th sentence, it can be seen that na is transliterated to near to the correct result whereas the seq2seq model could not transliterate ki accurately. Furthermore, it can be seen that the transformer model gives results as accurate as the widely used website for transliteration and also at times outperforms it.

B. RESULTS BY TRANSFORMER MODEL

Some of the accurate and erroneous results given by the transformer model are as follows:

Accurate Results		Erroneous Results	
yeh ek paycheeda jumla hai	یہ ایک پیچیدہ جملہ ہے	maujooda dor mein angraizi zaban ki ahmiyat barh gayi hai	موجود دور میں انگریزی زبان کی اہمیت بڑھ گئی ہے
inhen bachpan se hi ilm o adab aur drama se dilchaspi thi	انہیں بچپن سے ہی علم و ادب اور ڈرامہ سے دلچسپی تھی	woh chichawatni k rehnay walay hain	وہ چمچاوطنی کے رہنے والے ہیں
Iqbal musalmano ko phir isi akhuwat e islami ki taraf lotnay ki talqeen karte hain	اقبال مسلمانوں کو پھر اسی اخوت اسلامی کی طرف لوٹنے کی تلقین کرتے ہیں	iqbal musalmaanon ko phir isi akhuwat e islami ki taraf lotnay ki talqeen karte hain	کو پھر اسی احوت مسلمان اقبال اسلامی کی طرف لوٹنے کی تلقین کرتے ہیں تقوی صاحب پھر جنوری کو بمبئی کے راستے پانی کے جہادی سے کراچی آئے
Cheeni naib wazeer kharja ki wazeer e azam aur army cheif se mulaqaat	چینی نائب وزیر خارجہ کی وزیر اعظم اور آرمی چیف سے ملاقات	jo shakhs tujh say mangta hai usko day, taakay tujhe jo naahaq ka mil raha hai woh band nah hojaye	جو شخص تجھ سے مانگتا ہے اسکو دے, تاکہ تجھے جو ناہحق کا مل رہا ہے وہ بند نہ ہوجائے
angraizi ek bain ul aqwami zaban hai	انگریزی ایک بین الاقوامی زبان ہے	koh i noor roshni ka pahaar hai	کوہ کی نور روشنی کا پہیار ہے
taqreeban 70 feesad kaam hogaya hai	تقریبا 70 فیصد کام ہوگیا ہے	taqi sahab phir January ko Bombay ke rastay pani k jahaz se Karachi aye	تقوی صاحب پھر جنوری کو بمبئی کے راستے پانی کے جہادی سے کراچی آئے

Analysis of the Results:

The following findings can be made from the results given by the transformer model:

- **Complex words**

  Words such as paycheeda, naib, taqreeban which could possibly give multiple or inaccurate transliteration results are transliterated correctly.

- **Compound words**

  Words such as ilm o adab, wazeer e azam, akhuwat e islami are also transliterated correctly. This shows that the model has a good understanding of the words that are compound in Roman Urdu but have a two-word transliteration in Urdu script.

- **Long Sentences**

The transliteration result is not affected by the length of the sentence. It performs well on long sentences as well as short sentences.

- **Unknown words**

  As we can see in the 2nd erroneous sentence the word chichawatni is nowhere in the data set and is transliterated to a near to accurate result. The model tries its best to transliterate the word using one-to-one alphabet mapping of the word.

- **Incorrect Transliterations**

  The incorrect transliterations are near to correct and the model tries its best to give accurate results but due to not enough diversity in data it gives erroneous results.

7. CONCLUSIONS AND FUTURE WORK

This paper provides a comparison between the three models – RNN+LSTM, seq2seq, and transformer model to find the best model for Roman Urdu to Urdu transliteration purposes. After the comparison, we conclude that the Transformer model is the best model for the Roman Urdu-to-Urdu transliteration purpose. Further tuning of the Transformer model is performed and diversity in data is increased to get accurate, generalized, and according to context transliteration results that also address the rare word problem. As a result, the model gives us a BLEU score of around 75 which can be further improved by adding more diversity in the data set. In the future, we intend to build a web application for Roman Urdu to Urdu transliteration purposes and also increase the data set by collecting more data as well as adding diversity to it. Furthermore, the user interface for the application can also be changed and can be made more interactive by improving user experience.

REFERENCES

[1] Noureen, S. H. Huspi, and Z. Ali, “Sentiment Analysis on Roman Urdu Students’ Feedback Using Enhanced Word Embedding Technique,” *Baghdad Science Journal*, vol. 21, no. 2, pp. 725–739, 2024, doi: 10.21123/bsj.2024.9822.

[2] "Event Extraction Using Word Clustering and Word Embedding for Roman Urdu," *Journal of Hunan University Natural Sciences*, vol. 51, no. 1, 2024, doi: 10.55463/issn.1674-2974.51.1.13.

[3] J. A. Husain *et al.*, "RomanSetu: Efficiently unlocking multilingual capabilities of Large Language Models models via Romanization," Jan. 2024, [Online]. Available: http://arxiv.org/abs/2401.14280

[4] A. Khan, A. Ahmed, S. Jan, M. Bilal, and M. F. Zuhairi, "Abusive Language Detection in Urdu Text: Leveraging Deep Learning and Attention Mechanism," *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3370232.

[5] V. K. Chauhan, S. Singh, and A. Sharma, "HCR-Net: a deep learning based script independent handwritten character recognition network," *Multimed Tools Appl*, 2024, doi: 10.1007/s11042-024-18655-5.

[6] S. Khalid, C. Gao, G. Orynbek, and E. Tadesse, "Constructing a women-friendly academic ecology: understanding the push and pull forces on Pakistani women academics' research productivity," *Studies in Higher Education*, 2024, doi: 10.1080/03075079.2024.2322099.

[7] H. Raza and W. Shahzad, "End to End Urdu Abstractive Text Summarization with Dataset and Improvement in Evaluation Metric," *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3377463.

[8] R. Shahid, A. Wali, and M. Bashir, "Next word prediction for Urdu language using deep learning models," *Comput Speech Lang*, vol. 87, Aug. 2024, doi: 10.1016/j.csl.2024.101635.

[9] "Specifications Table", doi: 10.17632/d5j9fgbdcn.1.

[10] M. Arshad, B. Khan, K. Khan, A. M. Qamar, and R. U. Khan, "ABMRF: An Ensemble Model for Author Profiling Based on Stylistic Features Using Roman Urdu," *Intelligent Automation & Soft Computing*, vol. 0, no. 0, pp. 1–10, 2024, doi: 10.32604/iasc.2024.045402.

[11] F. Mehmood, H. Ghafoor, M. N. Asim, M. U. Ghani, W. Mahmood, and A. Dengel, "Passion-Net: a robust precise and explainable predictor for hate speech detection in Roman Urdu text," *Neural Comput Appl*, vol. 36, no. 6, pp. 3077–3100, Feb. 2024, doi: 10.1007/s00521-023-09169-6.

[12] S. Guha, “Empires, Languages, and Scripts in the Perso-Indian World,” *Comp Stud Soc Hist*, 2024, doi: 10.1017/S0010417523000439.

[13] A. S. Agrawal, B. Fazili, and P. Jyothi, “Translation Errors Significantly Impact Low-Resource Languages in Cross-Lingual Learning,” Feb. 2024, [Online]. Available: http://arxiv.org/abs/2402.02080

[14] M. Ayaz, S. Nizamani, A. A. Chandio, and K. K. Luhana, “Detection of Roman Urdu fraud/spam SMS in Pakistan Using Machine Learning,” *International Journal of Computing and Digital Systems*, vol. 15, no. 1, pp. 1053–1061, 2024, doi: 10.12785/ijcds/150174.

[15] T. Nasir and M. K. Malik, “Efficient CRNN: Towards end-to-end low resource Urdu text recognition using depthwise separable convolutions and gated recurrent units,” *Inf Process Manag*, vol. 61, no. 1, Jan. 2024, doi: 10.1016/j.ipm.2023.103544.

[16] S. Kanwal, M. K. Malik, Z. Nawaz, and K. Mehmood, “SEEUNRS: Semantically-Enriched-Entity-Based Urdu News Recommendation System,” *ACM Transactions on Asian and Low-Resource Language Information Processing*, Jan. 2024, doi: 10.1145/3639049.

[17] Y. A. Mohamed, A. Khanan, M. Bashir, A. H. H. M. Mohamed, M. A. E. Adiel, and M. A. Elsadig, “The Impact of Artificial Intelligence on Language Translation: A Review,” *IEEE Access*, vol. 12, pp. 25553–25579, 2024, doi: 10.1109/ACCESS.2024.3366802.

[18] T. Z. Shah, M. Imran, and S. M. Ismail, “A diachronic study determining syntactic and semantic features of Urdu-English neural machine translation,” *Heliyon*, vol. 10, no. 1, Jan. 2024, doi: 10.1016/j.heliyon.2023.e22883.

[19] K. Saifullah, M. I. Khan, S. Jamal, and I. H. Sarker, “Cyberbullying Text Identification based on Deep Learning and Transformer-based Language Models,” *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*, vol. 11, no. 1, Feb. 2024, doi: 10.4108/eetinis.v11i1.4703.

[20] N. Srivastava, G. Hinton, A. Krizhevsky, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” 2014.

[21] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: a Method for Automatic Evaluation of Machine Translation.”

[22] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.0473

[23] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.3215

[24] M.-T. Luong, I. Sutskever, Q. V. Le, O. Vinyals, and W. Zaremba, "Addressing the Rare Word Problem in Neural Machine Translation," Oct. 2014, [Online]. Available: http://arxiv.org/abs/1410.8206

[25] M. Alam and S. Ul Hussain, "Sequence to Sequence Networks for Roman-Urdu to Urdu Transliteration."

[26] Q. V. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," May 2014, [Online]. Available: http://arxiv.org/abs/1405.4053

[27] P. Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation." [Online]. Available: http://www.europarl.eu.int/

[28] S. Abid, A. Bukhari, P. Sajjad, A. Paracha, and P. D. Scholar, "Portrayal of Pakistan on Urdu Websites of BBC and VOA: A Framing and Audience Perception Analysis." [Online]. Available: http://xisdxjxsu.asia

[29] M. Haseeb, M. F. Manzoor, M. S. Farooq, U. Farooq, and A. Abid, "A versatile dataset for intrinsic plagiarism detection, text reuse analysis, and author clustering in Urdu," *Data Brief*, vol. 52, Feb. 2024, doi: 10.1016/j.dib.2023.109857.

[30] Y. Yu *et al.*, "Large Language Model as Attributed Training Data Generator: A Tale of Diversity and Bias." [Online]. Available: https://github.com/yueyu1030/AttrPrompt.

[31] X. Tan, T. Qin, J. Bian, T.-Y. Liu, and Y. Bengio, "Regeneration Learning: A Learning Paradigm for Data Generation," 2024. [Online]. Available: www.aaai.org

[32] H. Baruah, S. R. Singh, and P. Sarmah, "Transliteration Characteristics in Romanized Assamese Language Social Media Text and Machine Transliteration," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 2, Feb. 2024, doi: 10.1145/3639565.

[33] M. Khan and A. Srivastava, "Sentiment Analysis of Twitter Data Using Machine Learning Techniques," *International Journal of Engineering and Management Research Peer Reviewed & Refereed Journal e*, vol. 14, no. 1, 2024, doi: 10.5281/zenodo.10791485.

[34] S. Haque, Z. Eberhart, A. Bansal, and C. McMillan, "Semantic Similarity Metrics for Evaluating Source Code Summarization," in *IEEE International Conference on Program Comprehension*, IEEE Computer Society, 2022, pp. 36–47. doi: 10.1145/nnnnnnn.nnnnnnn.

[35] C. Shaib, J. Barrow, J. Sun, A. F. Siu, B. C. Wallace, and A. Nenkova, "Standardizing the Measurement of Text Diversity: A Tool and a Comparative Analysis of Scores," Mar. 2024, [Online]. Available: http://arxiv.org/abs/2403.00553

[36] T. Gaustad, C. A. McKellar, and M. J. Puttkammer, "Dataset for Siswati: Parallel textual data for English and Siswati and monolingual textual data for Siswati," *Data Brief*, p. 110325, Mar. 2024, doi: 10.1016/J.DIB.2024.110325.

[37] R. Van Der Goot, "Findings of the Association for Computational Linguistics Where are we Still Split on Tokenization?" [Online]. Available: https://github.com/machamp-nlp/

[38] O. Goldman, A. Caciularu, M. Eyal, K. Cao, I. Szpektor, and R. Tsarfaty, "Unpacking Tokenization: Evaluating Text Compression and its Correlation with Model Performance," Mar. 2024, [Online]. Available: http://arxiv.org/abs/2403.06265

[39] E. Lien Bolager, I. Burak, C. Datar, Q. Sun, and F. Dietrich, "Sampling weights of deep neural networks."

[40] G. Zaręba, M. Zarębski, and J. Smołka, "C++ and Kotlin performance on Android-a comparative analysis Analiza porównawcza wydajności języków C++ i Kotlin na platformie Android," 2024.

[41] T. Oh, S. Chung, B. Lunt, R. McMahon, and R. Rutherfoord, "The roles of IT education in IoT and data analytics," in *SIGITE 2017 - Proceedings of the 18th Annual Conference on Information Technology Education*, Association for Computing Machinery, Inc, Sep. 2017, pp. 39–40. doi: 10.1145/XXXXXXX.XXXXXXX.

[42] O. Zanevych, “ADVANCING WEB DEVELOPMENT: A COMPARATIVE ANALYSIS OF MODERN FRAMEWORKS FOR REST AND GRAPHQL BACK-END SERVICES,” *Grail of Science*, no. 37, pp. 216–228, Mar. 2024, doi: 10.36074/grail-of-science.15.03.2024.031.