https://doi.org/10.5564/jimdt.v6i1.3758

Joint NMT models for text conversion between traditional Mongolian script and cyrillic Mongolian: A comparative study

Uuganbaatar Dulamragchaa[1], Nomuundalai Batbileg[1*], Sainbayar Batbaatar[1], Purevsuren Tumurbaatar[2], Bilguutei Narantuya[1], Baatar Nyamkhuu[1], Tungalagtamir Bold[1], Bayarchimeg Enkhtaivan[2]

[1] *Institute of Mathematics and Digital Technology, Mongolian Academy of Sciences, Ulaanbaatar 13330, Mongolia*
[2] *Institute of Language and Literature, Mongolian Academy of Sciences, Ulaanbaatar 13330, Mongolia*

**Corresponding author: nomuundalai_ b@mas.ac.mn; ORCID:0009-0005-7259-4068*

Article Info: Received: 2024.09.15; Accepted: 2024.10.01; Published: 2024.12.26

Abstract: The research aims to develop a high-precision machine learning model for translating between Mongolian and Cyrillic scripts. Although Mongolian and Cyrillic scripts are different scripts of the Mongolian language, their sentence structure, number of words in a sentence, and position are the same. This article presents the results of our research on neural machine translation models tested for translating between the two scripts.

The Seq2Seq and Transformer models with attention mechanisms were trained using each tokenization method at the character, subword, and word levels. The accomplishments of these models were assessed regarding the conversion quality, computational efficiency, and the ability to handle the unique characteristics of the two scripts, and the benefits and drawbacks of each model were summarized.

It is believed that this comparison helps select the most suitable NMT model for similar tasks. The developed model has a vast potential for development and application as a system for translating between scripts, and it will facilitate some clerical work in public and private sector organizations at all levels providing services in Cyrillic and Mongolian scripts.

Key words: Machine Translation, Transformer, Neural machine translation(NMT), Deep Learning

1. Introduction

The Mongolian language is distinguished by being written in two different scripts: traditional Mongolian script and Cyrillic. In 1946, the Cyrillic script was officially used throughout Mongolia. From 2025, under the implementation of the "Law on the Mongolian Language"government organizations have conducted official documents in traditional Mongolian script and Cyrillic script, marking the beginning of becoming a "bi-script nation."In particular, Cyrillic script is a script with a history of about 80 years and is widely used in Mongolian daily communication, media, books, newspapers, magazines, and other social and economic spheres. At the same time, traditional Mongolian script has a history of several centuries of great cultural and historical importance. Since the two writing systems mentioned are phonetic scripts, they are the most sophisticated among the scripts used by humankind.

Since the Mongolian traditional and Cyrillic scripts are phonetic scripts expressing the characteristics of their agglutinating languages, they have many similarities and differences. In terms of writing style, traditional Mongolian script is written vertically, while Cyrillic script is written horizontally. Traditional Mongolian scripts' spelling rules do not change the form of word roots and suffixes, while Cyrillic scripts change the form of word roots and suffixes depending on the word structure when suffixes are added. On the other hand, the root of a word does not change in traditional Mongolian script, while the root of a word changes in Cyrillic script, and there are cases where a one-word root can be transformed into at least five forms. Furthermore, Mongolian homonyms are written separately in traditional Mongolian script, which is different from Cyrillic script in that they can be written separately in their grammatical form. However, since Mongolian is an agglutinating language, traditional Mongolian and Cyrillic scripts have the same sentence structure.

Thus, the duality of the writing system encounters a challenge in modern communication, documentation, and digital applications, mainly when it is necessary to translate one written text into another. Since manual translation is time-consuming, labor-intensive, and error-prone, there is a legitimate need to develop an automated solution that can reliably handle the complexity of two texts. Although this situation is an important and sought-after system to solve in connection with implementing the aforementioned law on conducting official service in "Dual Scripts,"there are often cases where traditional rule-based methods for translating text do not achieve high accuracy because of the significant differences in the structure and form of the two scripts. Neural machine translation translates text based on the context and general meaning of the sentence. It uses neural network models trained on parallel data sets. These models learn to predict the probability that a sequence of words in one language will lead to a sequence in another language, and the goal is to enable machines to produce simple and understandable translations. NMT models utilizing attention mechanisms and deep learning, including and Seq2Seq and Transformer, have made notable progress to carry out machine translation tasks. The significant benefits of these models are to learn from large data sets and predict based on the text's meaning, and the models are suitable for solving the difficulties of Mongolian script translation.

In this study, we aim to develop a high-accuracy system for converting text between traditional Mongolian and Cyrillic scripts using NMT models. In order to fulfill the aim, we intend to identify the optimal method for converting text by comparing different NMT architectures, Seq2Seq and Transformer, using different tokenization methods. In addition, this comparative study will examine the pros and cons of these models and determine which performs best in accuracy, speed, and computational efficiency. This study will not only contribute to the growing field of machine translation, but also have the significance of machine learning to recognize the differences between the two Mongolian writing systems. Based on our parallel sentence bank of Mongolian Cyrillic and traditional Mongolian script, we built this text-to-text conversion model using the artificial intelligence "Transformer" model. Moreover, we trained the Cyrillic-to-Mongolian script converter model and the Mongolian-to-Mongolian Cyrillic converter model separately based on the training data of the different features of the two scripts.

2. Methodology

2.1. Data processing

The data set for training the models was compiled from an open source such as github.com with team assistance, and the data set, which consists of parallel sentences written in Mongolian traditional and Cyrillic scripts, was prepared. The total data set contained approximately 105'110 parallel sentences.

Table 1: Data Set Information

Data Set	Description
Train set	94 599 (90%)
Validation set	10 511 (10%)
Test set	3 000 sentences

2.2. Utilized architectures

Seq2Seq with Attention Mechanism:

In the frame of translation between Cyrillic and Mongolian scripts, the Seq2Seq architecture with attention mechanism could help correctly handle the variance in style, typology, sentence structure, and word order of the two scripts.

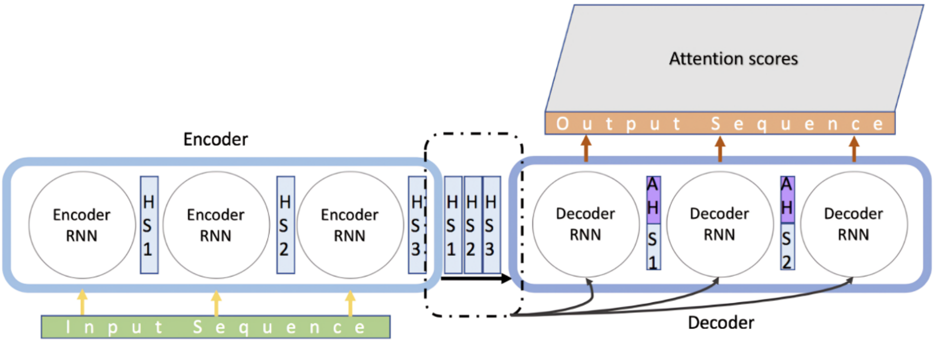


Figure 1: Seq2Seq architecture with an attention mechanism [1]

The architecture is composed of two major components:

- Encoder: processes the input sequence and creates a context vector.
- Decoder: generates an output sequence utilizing the context vector and attention mechanism.

The attention mechanism enables the decoder to concentrate on diverse parts of the input sequence at each output part. Attention scores between the current state of the decoder and each state of the encoder are calculated. These scores indicate how much attention the decoder should pay to each part of the input sequence.

Transformer:

The transformer architecture introduced in 2017 eliminated the need for recurrent layer (RNN, LSTM) and revolutionized sector of the native language processing (NLP). The main feature of the transformer architecture is based on the mechanism paying attention to itself, which allows the processing of all sequences at once. Thus, it accelerated the training and enhanced the quality of long data translation.

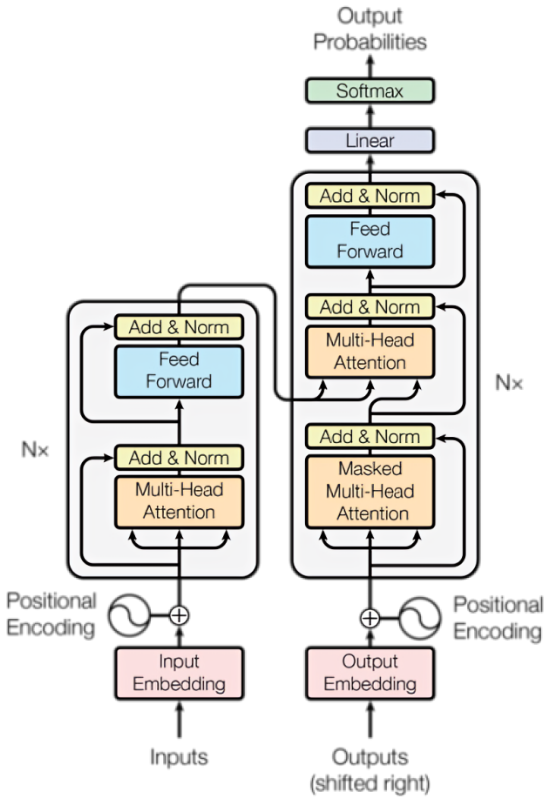


Figure 2: Transformer architecture [4]

1. **Input conversion**

   Token embedding:

   - Convert input tokens(word, sub-word, and character) into a dense vector of regular size.

   Positional encoding:

   - As the transformer does not define the position of each token, positional encoding is added to insert the token.
   - The sin function is used on an even index, while the cos function is applied on the d odd index to define the position of each token.

   $$\mathrm{PE}_{\mathrm{pos},\,2i} = \sin\left(\frac{\mathrm{pos}}{10000^{\frac{2i}{d_{\mathrm{model}}}}}\right), \mathrm{PE}_{\mathrm{pos},\,2i+1} \quad = \cos\left(\frac{\mathrm{pos}}{10000^{\frac{2i}{d_{\mathrm{model}}}}}\right). \tag{2.1}$$

2. **Encoder and Decoder**

   The transformer consists of two major parts.

   - Encoder: processes input sequence and generates high-level visual.
   - Decoder: creates output sequence based on Encoder visual and previously generated output.

   Each part is made of several similar layers.

3. **Encoder Layer**

   Each encoder layer contains the following items.

(a) Multi-Head and Self-Attention Mechanism

- It determines the correlation between tokens through input sequence.
- It calculates attention score using Scaled dot-product:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{2.2}$$

- Matrices such as Q (Query), K (Key), and V (Value) are sets of vectors learned from the input embeddings.
- d_k represents the dimensionality of the key vectors.

The model is multi-head to enable you to focus on different parts of the input:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O. \tag{2.3}$$

- W^O is a matrix of learned projection.

(b) Feed-Forward Neural Network (FFN)

- It applies a fully connected network on each position.
- It is made up of a bilinear transform with ReLU activation.

$$\text{FFN}(x) = \max(0,\, xW_1 + b_1)W_2 + b_2. \tag{2.4}$$

(c) Add Norm

$$\text{Output} = \text{LayerNorm}(x + \text{sublayer}(x)). \tag{2.5}$$

The Transformer architecture is used due to the following advantages:

- **Parallelism:** Unlike the Seq2Seq architecture processing sequence in one step, transformer architecture processes all sequences in parallel and hastens the training process and inference time.
- **Working with long sequences:** The self-attention mechanism helps the transformer learn long sequences efficiently, which is suitable for processing long sentences and paragraphs while transforming the script.

Comparison of Seq2Seq and Transformer architectures:

Regarding converting between traditional Mongolian and Cyrillic scripts, Seq2Seq and Transformer architectures have their benefits. Seq2Seq architecture is more effective for simple sentence structures with short sequences, and it does not perform well on more complex long sentence structures since it relies on a single context vector. Therefore, Transformer architecture is efficient in working with long sentences and determining the relationships of complex text structures because it has an attention mechanism to overcome this disadvantage.

2.3. Tokenization methods

“Character level” or word- level tokenization:

Word-level tokenization tokenizes a word of input text by dividing it as a unit. For instance, the sentence written in the Cyrillic script, such as "I am reading a book,"is tokenized into word levels [I, a book, reading, and am]. The tokenization method is consistent with human explanation. However, it is difficult to process a morphologically rich language like Mongolian, which has a large number of word forms due to inversions, derivatives, and compound words.

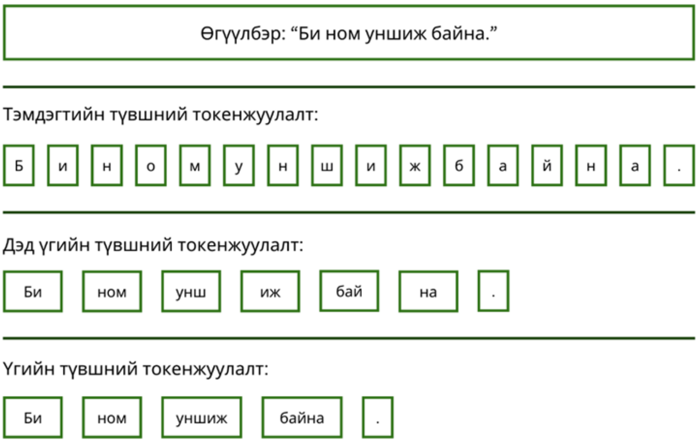

Өгүүлбэр: “Би ном уншиж байна.”
Тэмдэгтийн түвшний токенжуулалт:
Б
и
н
о
м
у
н
ш
и
ж
б
а
й
н
а
.
Дэд үгийн түвшний токенжуулалт:
Би
ном
унш
иж
бай
на
.
Үгийн түвшний токенжуулалт:
Би
ном
уншиж
байна
.

Figure 3: Tokenization methods

“Sub-word level” or sub-word level tokenization:

The main advantage of sub-word level tokenization is interpolating between tokenizations based on word and character. For unknown words, this enables to tokenize characters after root words based on the words in the vocabulary. For example, a sentence like "холоо одсон хүний хань"will be tokenized as follows: ["хол", "оо", "од", "сон", "хүн", "ий", "хан", "ь"].

“Word-level” or character-level tokenization:

Character-level tokenization operates by treating a character of each word of input text word as a unit. For instance, a word written in a traditional script like "ᠠᠪᠠ" (means father) is tokenized as ["ᠠ," "ᠪ," "ᠠ"]. The character-level tokenization allows the model to learn at most minor levels and is suitable for working with rare or unknown words.

Comparison of tokenization-based models:

1. **Transformer with character-level tokenization:** The transformer model with character-level tokenization showed high performance while processing the rule of traditional Mongolian script. The self-attention mechanism enables the model to save context at the character level and translate accurately a text with common and rare words. However, a rise in the sequence length led to an increase in estimation cost, and it is inappropriate to use character-level tokenization.

2. **Transfomer with sub-word-level tokenization:** The sub-word-level tokenization kept a balance between vocabulary size and efficiency. The attentiveness of the transformer model applying a multi-head attention mechanism was longer on sub-word relationships and practical in processing extended and more complex sentences. The model could integrate different word forms, making it suitable for large-scale text translation tasks.

3. **Seq2Seq with word-level tokenization:** Due to work efficiency in a short sequence and training simplicity, the Seq2Seq model was combined with word-level tokenization. Although the Seq2Seq model with attention mechanism is practical for processing short, it could not fully convert words which are not included in vocabulary and long sentences when using word-level tokenization. One limitation of using word-level tokenization in the Seq2Seq model is the large vocabulary size caused by the morphological complexity of the Mongolian language. It leads to missing words when drawing inferences, and the model encounters words and word forms that were not learned during training.

Conclusion on tokenization:

Through this comparison, the tokenization option affects the performance of each model. With its feature and sub-word level tokenizer, the transformer model can overcome the complex differences in style and spelling between traditional and Cyrillic Mongolian scripts. Consequently, integrating the transformer model with sub-word or character-level tokenization produced reliable and accurate results regarding complex tasks, including text conversion between traditional and Cyrillic Mongolian scripts.

2.4. Model evaluation

After training, we must evaluate the model's translation on pre-trained test data."BLEU" (Bilingual Evaluation Understudy) is a benchmark for assessing machine translation quality. However, it is considered inappropriate to evaluate tokenized translation task at character-level for several reasons.

1. **Word-level loss:** BLUE is designed to work at the word or n-gramm level. If the basic unit is a character in character-level tokenized translation, word-level loss of BLUE leads to incorrect evaluation.
2. **Lack of linguistic information:** BLEU does not contain information on language knowledge and basic structure. Character-level tokenized translation focuses on characters instead of words and phrases, while issues such as morphology, syntax, or semantics may be related to assessing translation quality.
3. **Insensitivity to structural changes:** BLEU is not responsive to structural changes in the translation. Tokenized translation at the character level frequently embodies significant structural changes, including changes in word order and sentence structure, compared with word-level translation. The BLUE cannot adequately maintain the quality of structural change.

For these reasons, researchers utilize alternative assessment metrics or formulate evaluation methods designated for tokenized translation tasks at the character level. The metrics can concentrate on character-level similarity and criteria related to language features or tasks to evaluate translation quality precisely. "Word Error Rate" (WER) is a widely used metric for assessing the accuracy of language translation methods. "Word Error Rate" (WER) measures the variance between predicted translation output and targeted translation by a quantitive index. The metric serves as a standard for evaluating the accuracy of translated text and models' performances.

$$\text{WER} = \frac{S_w + D_w + I_w}{N_w}. \tag{2.6}$$

3. The results

The models' performances were evaluated through Word Error Rate (WER). The WER is a standard metric used to measure the difference between predicted and actual word order.

Table 2: Comparison of models' WER scores

Model + Tokenization	**WER% (Cyrillic → Mongolian Script)**	**WER% (Mongolian Script → Cyrillic)**
Transformer + Char	11.0%	8.0%
Transformer + Subword	15.7%	–
Seq2Seq + Attn + Word	37.1%	36.2%

"Cross Entropy Loss" is a standard choice of natural language processing models such as machine translation. It measures the variance between the predicted probability distribution and target distribution and helps us optimize our model parameters.

$$\mathrm{LCE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J} y_{ij}\log\left(P_{ij}\right). \tag{3.1}$$

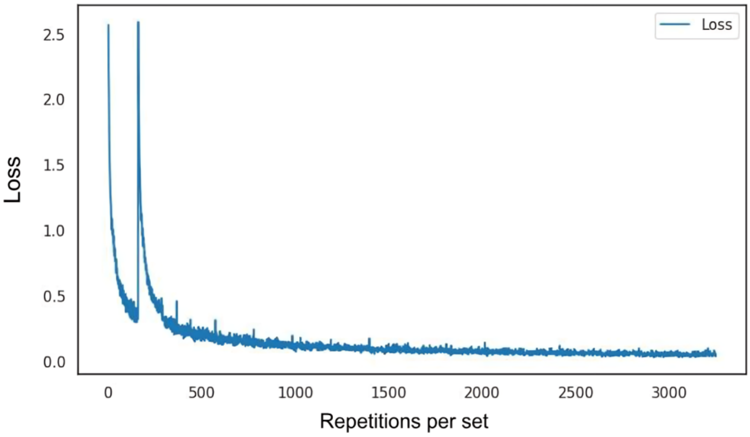


Figure 4: Loss graph of the "Transformer + Char" model for converting from Cyrillic to Mongolian traditional script

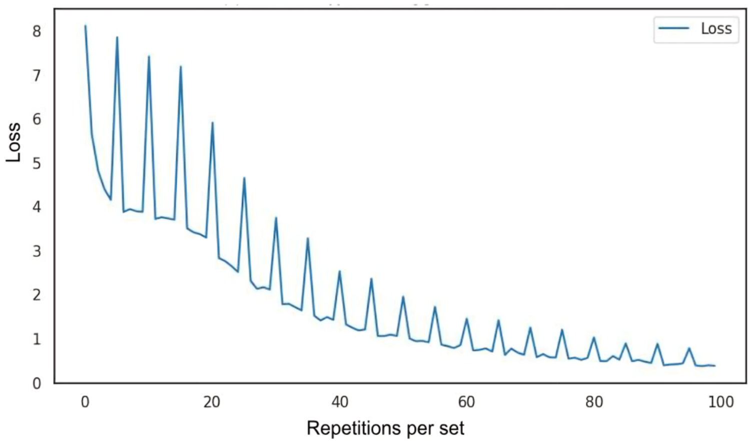
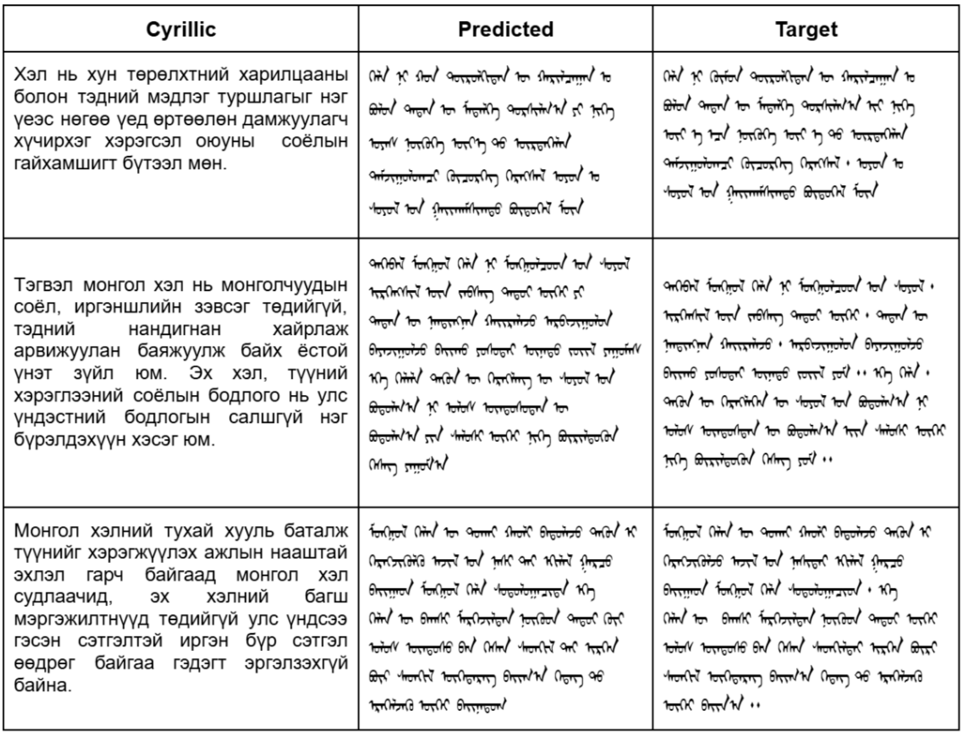


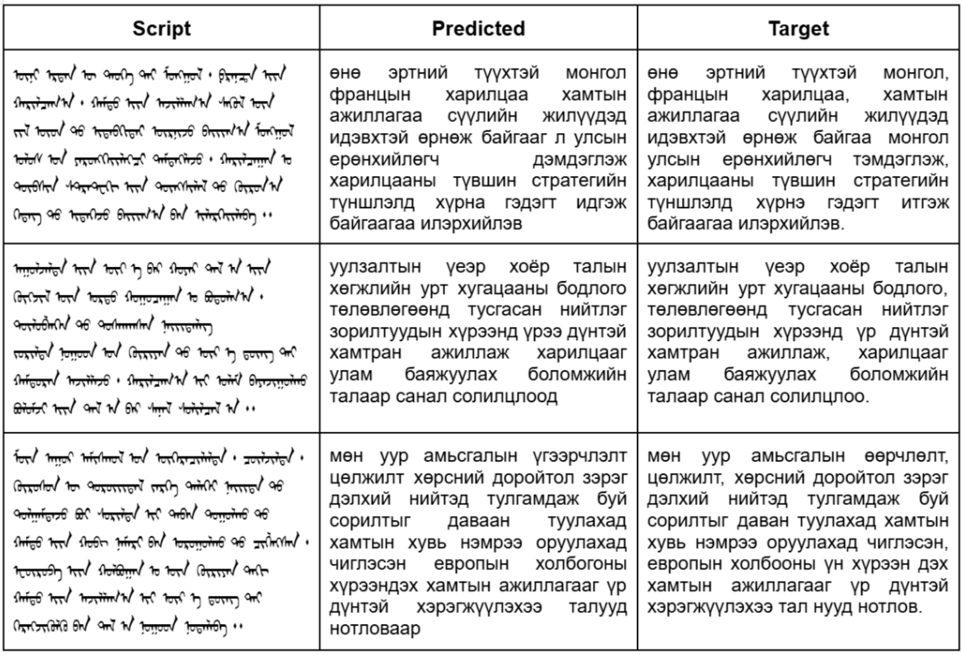
Figure 5: Loss graph of the "Transformer+Char" model for converting from Mongolian traditional script to Cyrillic script

The result of the Transformer model for converting between Mongolian and Traditional Mongolian script with character level tokenizer is shown in the following two figures.



Cyrillic	Predicted	Target
Хэл нь хун төрөлхтний харилцааны болон тэдний мэдлэг туршлагыг нэг үеэс нөгөө үед өртөөлөн дамжуулагч хүчирхэг хэрэгсэл оюуны соёлын гайхамшигт бүтээл мөн.	[illegible]	[illegible]
Тэгвэл монгол хэл нь монголчуудын соёл, иргэншлийн зэвсэг төдийгүй, тэдний нандигнан хайрлаж арвижуулан баяжуулж байх ёстой үнэт зүйл юм. Эх хэл, түүний хэрэглээний соёлын бодлого нь улс үндэстний бодлогын салшгүй нэг бүрэлдэхүүн хэсэг юм.	[illegible]	[illegible]
Монгол хэлний тухай хууль баталж түүнийг хэрэгжүүлэх ажлын наaштай эхлэл гарч байгаад монгол хэл судлаачид, эх хэлний багш мэргэжилтнүүд төдийгүй улс үндсээ гэсэн сэтгэлтэй иргэн бүр сэтгэл өөдрөг байгаа гэдэгт эргэлзэхгүй байна.	[illegible]	[illegible]

Figure 6: Results of converting from Cyrillic Mongolian to Traditional Mongolian script (Transformer + Char)



Script	Predicted	Target
[illegible]	өнө эртний түүхтэй монгол францын харилцаа хамтын ажиллагаа сүүлийн жилүүдэд идэвхтэй өрнөж байгааг л улсын ерөнхийлөгч дэмдэглэж харилцааны түвшин стратегийн түншлэлд хүрна гэдэгт идгэж байгаагаа илэрхийлэв	өнө эртний түүхтэй монгол, францын харилцаа, хамтын ажиллагаа сүүлийн жилүүдэд идэвхтэй өрнөж байгаа монгол улсын ерөнхийлөгч тэмдэглэж, харилцааны түвшин стратегийн түншлэлд хүрнэ гэдэгт итгэж байгаагаа илэрхийлэв.
[illegible]	уулзалтын үеэр хоёр талын хөгжлийн урт хугацааны бодлого төлөвлөгөөнд тусгасан нийтлэг зорилтуудын хүрээнд үрээ дүнтэй хамтран ажиллаж харилцааг улам баяжуулах боломжийн талаар санал солилцлоод	уулзалтын үеэр хоёр талын хөгжлийн урт хугацааны бодлого, төлөвлөгөөнд тусгасан нийтлэг зорилтуудын хүрээнд үр дүнтэй хамтран ажиллаж, харилцааг улам баяжуулах боломжийн талаар санал солилцлоо.
[illegible]	мөн уур амьсгалын үгээрчлэлт цөлжилт хөрсний доройтол зэрэг дэлхий нийтэд тулгамдаж буй сорилтыг даваан туулахад хамтын хувь нэмрээ оруулахад чиглэсэн европын холбогоны хүрээндэх хамтын ажиллагааг үр дүнтэй хэрэгжүүлэхээ талууд нотловаар	мөн уур амьсгалын өөрчлөлт, цөлжилт, хөрсний доройтол зэрэг дэлхий нийтэд тулгамдаж буй сорилтыг даван туулахад хамтын хувь нэмрээ оруулахад чиглэсэн, европын холбооны үн хүрээн дэх хамтын ажиллагааг үр дүнтэй хэрэгжүүлэхээ тал нууд нотлов.

Figure 7: Part of the result of converting from traditional Mongolian script to Cyrillic Mongolian (Transformer + Char)

4. Conclusion

In the frame of this research, we compared neural machine translation models based on the "Seq2Seq with Attention" and "Transformer" architectures to build an artificial intelligence model for converting between traditional Mongolian and Cyrillic texts and determined the most effective converter model. The results of the study and experiments show that the model combining the Transformer architecture with character-level tokenization performed best in terms of its ability to accurately consider differences in the form, structure, and grammar of texts under limited data conditions. When evaluating the model on test data, the WER for converting from Cyrillic to Mongolian script was 12.5% and BLEU was 32.36%, while the WER for converting from Mongolian script to Cyrillic was 7.9% and BLEU: 28.05%.

From the result comparing some NMT models, the model combining the Transformer architecture and character-level tokenization is the most optimal solution for our complex task of converting between Mongolian script and Cyrillic texts. In the future, the quality of the translation can be improved by expanding the dataset, improving the tokenization method to be compatible with language features, and testing new architectures and technologies.

References

[1] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence-to-Sequence Learning with Neural Networks," *Advances in Neural Information Processing Systems*, Vol. 27, Curran Associates Inc., 2014.

[2] *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Curran Associates Inc., 2014.

[3] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *arXiv preprint arXiv:1409.0473*, 2015.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett, and D. Precup, Curran Associates Inc., 2017, pp. 5998–6008.

[5] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1715–1725, 2016.

[6] A. Mao, et al., "Cross-Entropy Loss Functions: Theoretical Analysis and Applications," In *International Conference on Machine Learning*, 2023, pp. 23803–23828.