JIMDT, Vol. 6, 2024
RESEARCH ARTICLE

https://doi.org/10.5564/jimdt.v6i1.3599

Cyrillic Mongolian-to-Traditional Mongolian Conversion Method Based on the Transformer

Muhan Na[1,2,3*], Feilong Bao[1,2,3],Weihua Wang[1,2,3],Guanglai Gao[1,2,3],Uuganbaatar Dulamragchaa[4]

[1] *College of Computer Science, Hohhot, Inner Mongolia University 010021, China,*
[2] *Inner Mongolia Key Laboratory of Mongolian Information Processing Technology, Hohhot, Inner Mongolia University 010021, China*
[3] *National Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian, Hohhot, Inner Mongolia University 010021, China*
[4] *Institute of Mathematics and Digital Technology, Mongolian Academy of Sciences, Ulaanbaatar, Mongolia*

**Corresponding author: Namuhan_NMH@163.com*

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

Abstract: Cyrillic Mongolian and Traditional Mongolian are primarily utilized in Mongolia and China. The task of converting Cyrillic Mongolian to Traditional Mongolian (C2T) plays a vital role in facilitating language communication between compatriots of both nations and holds significant importance in the scientific, economic, and cultural domains of both countries. Mongolian words consist of stems and suffixes, resulting in an extensive Mongolian vocabulary that includes a multitude of Out-of-vocabulary (OOV) words. The conversion of OOV words cannot be effectively addressed solely through the use of rules and dictionaries. Hence, this paper presents a Transformer-based approach for Cyrillic Mongolian to Traditional Mongolian conversion. Experimental results demonstrate a 5.72% reduction in word error rate (WER) compared to the joint sequence approach.

Key words: Neural Network; Self-Attention; Mongolian translation

1. Introduction

Mongolian, which is part of the Mongolic language family, is classified within the Altaic language family and is notably the largest and most prominent member in this language group [1]. It holds a unique position as both the most widely spoken and most recognized language within the Mongolic family. In Mongolia, the Mongolian script is currently written using Cyrillic characters, while in Inner Mongolia, China, the Traditional Mongolian script is used [2]. The task of converting Cyrillic Mongolian to Traditional Mongolian (C2T) aims to transform text written in the Cyrillic Mongolian script into its equivalent Traditional Mongolian script. Cyrillic Mongolian and Traditional Mongolian are agglutinative languages, meaning they form words by adding multiple

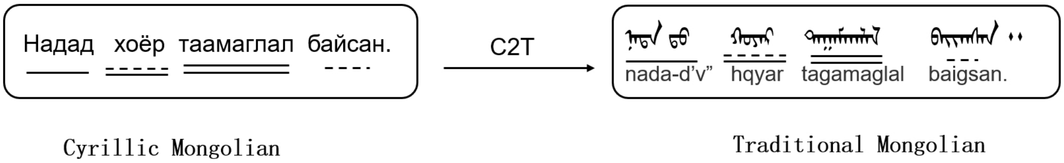
Надад хоёр таамаглал байсан.
C2T
nada-d'v" hqyar tagamaglal baigsan.
Cyrillic Mongolian
Traditional Mongolian

Figure 1: Example of C2T Conversion

suffixes to the stem. This process generates a vast vocabulary that cannot be completely captured in a dictionary. Furthermore, the distinct word formation rules of the two scripts create a one-to-many correspondence challenge, making it challenging to establish a direct mapping between Cyrillic Mongolian and Traditional Mongolian. Figure 1 provides an example of C2T conversion, Cyrillic Mongolian and Traditional Mongolian corresponding words is represented by the same underline.

The conventional approach primarily focuses on utilizing rule-based and statistical models, such as the joint sequence model, to tackle the C2T task. Notably, Bao et al.**[3]** employ the joint sequence model for the C2T task. Later, in 2017, Bao et al. **[4]** introduce a hybrid approach that combines rule-based and joint sequence models. This approach utilizes rule-based techniques to convert words within the existing vocabulary and utilizes the joint sequence model for converting out-of-vocabulary (OOV) words. However, both rule-based and statistical conversion methods have their limitations. Firstly, the rule-based approach falls short when it comes to handling out-of-vocabulary (OOV) words and loanwords. Mongolian words consist of stems and suffixes, and different types of suffixes can be attached to the same stem, resulting in a vast vocabulary. Additionally, the Mongolian language incorporates numerous loanwords that do not adhere to the regular word-formation rules. Secondly, statistical models have limited generalization ability and modeling capacity. The joint sequence model, for instance, utilizes a shallow architecture and lacks robust nonlinear modeling capabilities **[5,6]**. The C2T task can be considered as a typical machine translation task between two languages, making it fall into the category of sequence-to-sequence (Seq2Seq) modeling. The "encoder-decoder"architecture has shown successful applications in various Seq2Seq tasks, including neural machine translation **[5,7,8]**, speech synthesis **[9,10,11,12,13]**, grapheme to-phoneme (G2P) conversion **[14,15,16]**, and more. Recently, significant research efforts have been focused on recurrent neural networks (RNN), self attention-based "encoder-decoder"models , and pre-training language models. Although there is limited data available for training large-scale pre-training models specifically for the C2T task, both RNN-based encoder-decoder models and self-attention models have displayed impressive performance in Seq2Seq tasks. In this paper, we propose a conversion method that based on the Transformer, Compared with the previous method, performance has been improved to a certain extent.

2. Task Challenges

2.1. Scarce resources

Having a sizable corpus of aligned sentence-level data is crucial for machine translation tasks. However, this paper tackles the C2T task at the word-level, recognizing the challenge posed by the scarcity of abundant and high-quality training data. The low-resource nature of the Mongolian language contributes to the current lack of such data.

2.2. Agglutinative Features

Cyrillic Mongolian and Traditional Mongolian exhibit similarities in terms of word formation, pronunciation, and grammar rules. However, they diverge in their symbol systems, morphological rules, and the relationship between pronunciation and spelling.

Symbol systems: The Cyrillic Mongolian script comprises 13 vowels, 20 consonants, 1 hardened character, and 1 softened character **[4]**. In contrast, Traditional Mongolian consists of 8 vowels and 27 consonants [4]. Cyrillic Mongolian differentiates between uppercase and lowercase letters, with specific rules for capitalization of the initial letter, while Traditional Mongolian does not follow such rules. Additionally, Traditional Mongolian exhibits inconsistencies between its code and presentation form. To address this issue, the present study focuses on converting Traditional Mongolian characters into their corresponding Latin transcriptions **[17]**.

Morphological rules: The morphological rules in Cyrillic Mongolian encompass 66 categories **[1]**, whereas Traditional Mongolian consists of only 4 categories **[18, 19]**. This fundamental difference means that there is no direct one-to-one correspondence between the characters of the two Mongolian scripts.

Correlation between pronunciation: The pronunciation and spelling of Cyrillic Mongolian exhibit a direct one-to-one correspondence. Conversely, Traditional Mongolian demonstrates a one-to-many relationship. This results in a Cyrillic Mongolian word corresponding to multiple traditional Mongolian words when converting from Cyrillic Mongolian to traditional Mongolian. The unique characteristics of word formation in Mongolian contribute to its exten sive lexicon and a significant number of unfamiliar words. Additionally, the diverse morphological rules make it challenging for models to learn the corre spondence between two Mongolian characters. The issues of data sparsity and the agglutinative nature of Mongolia pose substantial obstacles for the C2T task at hand. In the following section, we will explore the linguistic knowledge described above to successfully carry out the C2T conversion.

3. Previous method

The C2T conversion process is divided into three steps: Cyrillic Mongolian text preprocessing, using a rule based approach to convert in-vocabulary words. Moreover, using a Joint sequence model to convert words that rules can‘t successfully convert; finally, when a Cyrillic Mongolian word corresponds to multiple Traditional Mongolian words, a language model is used to select the optimal sequence. The C2T conversion process combining rules and Joint sequence model is shown in Figure 2.

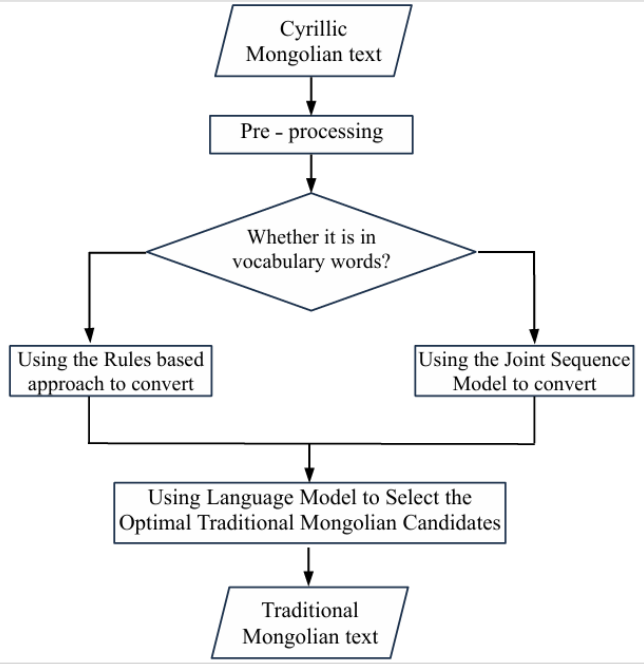
Cyrillic Mongolian text
Pre - processing
Whether it is in vocabulary words?
Using the Rules based approach to convert
Using the Joint Sequence Model to convert
Using Language Model to Select the Optimal Traditional Mongolian Candidates
Traditional Mongolian text

Figure 2: C2T conversion process combining rules and Joint Sequence Model

3.1. Preprocessing

Firstly, we segment and tokenize the Cyrillic Mongolian text. Secondly, we identify and preserve non-Traditional Mongolian characters while converting the regular punctuation and numerical symbols. Thirdly, the abbreviations of Cyrillic Mongolian are recognized and restored based on the abbreviation dictionary.

3.2. Rule based T2C Conversion

Mongolian words consist of stems and suffixes, with suffixes further cat egorized into derivational suffixes and inflectional suffixes. When a stem is Figure 3 C2T conversion based on Transformer combined with an inflectional suffix, the semantics of the word remain the same while the grammatical meaning changes, such as verb tense, aspect, and mood. Conversely, when a stem is combined with a derivational suffix, the semantics change, creating a new word. This paper focuses exclusively on inflectional suf fixes. The rule-based C2T conversion can be divided into three steps: 1. Split the Cyrillic Mongolian words into stems and suffixes. 2. Match the correspond ing Traditional Mongolian text from the stem and suffix libraries. 3. According to Traditional Mongolian word formation rules, splice the stems and suffixes into the Cyrillic Mongolian word.

The segmentation of suffixes in Cyrillic Mongolian is a more intricate pro cess. When forming the Cyrillic Mongolian words, vowels and consonants are dropped, generated, and transformed. Therefore, during suffix segmentation of Cyrillic Mongolian words, there are procedures involved in vowel and con sonant recovery, loss, and reduction. Considering the characteristics of word formation in Cyrillic Mongolian, this paper has compiled over 30 rules for suf f ix segmentation based on references such as the "Cyrillic Mongolian Learning Book"[20] and "Mongolian Grammar"**[1]**. Additionally, referencing "Mongo lian Grammar"**[1]** and "Modern Mongolian"**[21]**, this article has summarized over 20 rules for word stem and suffix concatenation in Traditional Mongolian. We establishes a stem library (containing 63,501 entries), a verb suffix com parison library (containing 1,626 entries), and a static word suffix comparison library (containing 694 entries) based on "Mongolian Dictionary"**[22]**, "New Mongolian Chinese Dictionary"**[23]**, and "A comparative study of Mongolian and Cyrillic orthography"**[2]** .

3.3. C2T conversion based on Joint sequence model

For words that cannot be converted by the rules, the joint sequence model is used for conversion. The idea of the joint sequence model is to represent the relationship between the input sequence and the output sequence through a series of joint units consisting of common sequences of input and output symbols. During decoding, an N-gram language model is used to predict characters. The best model configuration can be found by adjusting the value of N in N-gram.

3.4. Optimal Sequence Selection

During the C2T conversion process, one Cyrillic Mongolian word corresponding to multiple Traditional Mongolian words. Therefore, this paper uses an N-gram language model to select the optimal sequence that fits the context. After completing the above steps, the entire T2C conversion is completed.

4. C2T conversion based on Transformer

Cyrillic Mongolian and Traditional Mongolian both belong to phonetic scripts, where each word in Mongolian is represented as a sequence of letters. Therefore, the C2T conversion task can be viewed as a Sequence-to-Sequence (Seq2Seq) task. Recurrent Neural Networks (RNN) and Transformer have been extensively used in Seq2Seq tasks and have demonstrated remarkable performance. In this paper, the Transformer architecture is employed for converting Cyrillic Mongolian to Traditional Mongolian. Figure 3 illustrates the three main components of the C2T conversion: "Text Pre-processing,Encoder-Decoder,"and "Post-processing."In the Text Pre-processing stage, the Cyrillic Mongolian word is taken as input, and a character sequence is generated as output. The Encoder then reads the character sequence and generates a high-level representation, which is subsequently passed to the Decoder. The Decoder utilizes the hidden representation to predict the Latin transcriptions of Traditional Mongolian characters. Finally, the Post-processing module converts the Latin transcriptions into the Traditional Mongolian script, yielding the final conversion results.

To verify the effectiveness of encoder-decoder structures based on RNN and Transformer for the T2C task, this paper conducted experiments on the RNN encoder-decoder structure without attention (RNN), the RNN encoder-decoder structure with attention (RNN+ATT), and the Transformer encoder-decoder structure. Here, we will introduce the internal structures of the RNN and Transformer encoder-decoder models. In the RNN-based encoder-decoder structure, the encoder part consists of multiple layers of Bidirectional long short-term memory (BiLSTM) networks, which convert the input Cyrillic Mongolian sequence C into a high-dimensional vector representation h. The decoder uses h and the previously predicted outputs $o_1, o_2, \ldots, o_{(t-1)}$ as inputs to calculate the probability distribution of the next word o_t. The posterior distribution of the decoder's predicted output at time step h is generated from the decoder state s_t. Within the attention-based RNN encoder-decoder framework, the predicted output's posterior distribution at time step h is formulated by merging the decoder state s_t with the context vector a_t. The context vector a_t is computed using an attention module that takes into account the hidden states of both the encoder and decoder. In contrast to RNN, the Transformer encoder-decoder structure uses self-attention mechanisms to build the entire model framework. In the Transformer architecture, the encoder is constructed with N network blocks, incorporating multi-head attention layers and fully connected feed forward neural network layers. Likewise, the decoder is composed of N network blocks, comprising multi-head attention layers, masked multi-head attention layers, and fully connected feed forward neural network layers. Residual connections are utilized between each layer to enhance network optimization, with the addition of normalization.

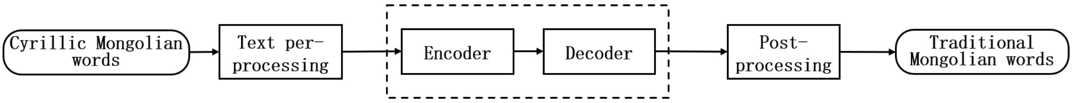

Figure 3: C2 conversion based on tRansformer

5. Experiments and Conclusions

5.1. ExperimentalData

This paper collected 63,668 word pairs of Cyrillic Mongolian words and Traditional Mongolian words from the "New Mongolian Chinese Dictionary"**[23]**, denoted as “Mon_data”. Among them, 58,436 word pairs are randomly selected as the training set and 5,232 word pairs are selected as the test set. 3.2 million Traditional Mongolian sentences are led from Traditional Mongolian news websites for training the Traditional Mongolian language model. To test the method proposed in this paper, 8,992 sentences are collected and organized as the "T2C_sentence"
test set, containing 99,010 words, including 93,546 in-vocabulary words and 5,464 OOV words.

5.2. Evaluation Metrics

The evaluation metrics used in this paper are Word Error Rate (WER) and Character Error Rate (CER), which are calculated by Equations as follows, respectively.

$$WER = 1 - \frac{N_{corrent}}{N_{total}} \tag{5.1}$$

$$CER = \frac{N_{ins} + N_{del} + N_{sub}}{N_{charater-total}} \tag{5.2}$$

The $N_{current}$ is the number of correctly converted Mongolian words, N_{total} is the total number of Mongolian words to be converted, N_{ins} is the number of letter insertion errors during conversion, N_{del} is the number of letter deletion errors during conversion, N_{sub} is the number of letter substitution errors during conversion, and $N_{character-total}$ is the total number of letters in the word.

5.3. Experimental Results of Transformer

This paper compares the performance of the Joint sequence model, RNN-based encoder-decoder model, and Transformer model on the C2T task. The Joint sequence model uses a joint unit consisting of input and output symbols to represent the relationship between the input sequence and the output sequence using a common sequence. During decoding, an N-gram language model is used for character prediction. This paper will search for the optimal configuration of the Joint sequence model by adjusting the order of the N-gram.

	WER%	CER%
Joint (Num=1)	89.79	26.39
Joint (Num=2)	70.53	16.47
Joint (Num=3)	50.34	10.13
Joint (Num=4)	33.75	6.51
Joint (Num=5)	26.19	4.88
Joint (Num=6)	23.24	4.34
Joint (Num=7)	22.76	4.22
Joint (Num=8)	22.69	4.21
Joint (Num=9)	22.63	4.2
Joint (Num=10)	22.67	4.2

Table 1: C2T conversion results based on the joint sequence

Two RNN-based architectures are implemented in this paper: RNN encoder-decoder structure without attention mechanism (short as "RNN") and RNN encoder-decoder structure based on attention mechanism (short as "RNN+ATT"). The RNN hidden layer size is set to 512 and 1024, and the number of hidden layers is set to 1, 2, and 4 to explore the optimal model configuration. During model training, the input to the encoder is a 128-dimensional character sequence, and the decoder output is Cyrillic Mongolian characters. The RNN parameters are set: batch size=32, epochs=100,learning rate=0.0005. The learning rate is reduced by a factor of 0.9 every 20 epochs. The cross-entropy loss function is used.

The "Transformer-Tiny"model**[21]** is used to construct the Transformer-based encoder-decoder model. Compared with the "Transformer-base"model, the "Transformer-Tiny"is more suitable for tasks with smaller vocabularies, making it more suitable for character level C2T conversion tasks. In order to find the best parameter combination, the number of attention mechanism heads is tested at 2 and 4, while the number of hidden layers is tested at 1, 2, 4, and 6. All comparative experiments are performed using Tensor2Tensor. In the experiments, the input dimension is fixed at 128, the training proceeds for 100,000 steps, and each batch consists of 4096 samples. The learning rate is set to 0.2, and there are 8000 warm-up steps according to the approach described in reference**[16]**. All word embeddings are initialized randomly and updated across the entire model.The C2T conversion experiments reveal that augmenting the order of the N-gram language model from 1 to 10 leads to a reduction in both WER and CER. The best performing joint sequence model had an N-gram language model order of 9, denoted as Joint (Num=9), with a WER of 22.63% and a CER of 4.2%, as shown in Table 1.

Table 2: Results of C2T conversion based on RNN

	WER%	CER%
RNN (512U_1L)	23.15	4.82
RNN (512U_2L)	25.94	5.63
RNN (512U_4L)	29.07	7.26
RNN (1024U_1L)	22.46	4.56
RNN (1024U_2L)	24.14	5.04
RNN (1024U_4L)	27.81	7
RNN+ATT (512U_1L)	20.95	4.11
RNN+ATT (512U_2L)	21.58	4.22
RNN+ATT (512U_4L)	20.05	3.93
RNN+ATT (1024U_1L)	21.12	4.04
RNN+ATT (1024U_2L)	19.51	3.72
RNN+ATT (1024U_4L)	21.85	4.14

Table 3: Results of C2T conversion based on the Transformer

	WER%	CER%
Trans (1S_2H)	25.5	4.95
Trans (2S_2H)	19.25	3.59
Trans (4S_2H)	17.32	3.29
Trans (6S_2H)	17.14	3.2
Trans (1S_4H)	22.87	4.44
Trans (2S_4H)	17.99	3.36
Trans (4S_4H)	17.07	3.18
Trans (6S_4H)	16.91	3.15

According to the Table 2, the best performance is achieved by the 1-layer Bidirectional LSTM (BiLSTM) with 1024 hidden units. This model achieves a Character Error Rate (CER) of 4.56% and a Word Error Rate (WER) of 22.46%. Interestingly, the optimal result, RNN(1024U_1L), is comparable to the performance achieved with N=9.

To further analyze the impact of hidden unit sizes, we keep the number of hidden layers fixed in the BiLSTM model and adjust the hidden unit size to observe the performance. It is observed that increasing the hidden unit size can improve the model's performance. For instance, the RNN(512U_1L) model outperforms the RNN(1024U_1L) model. However, stacking more hidden layers does not always lead to significant improvements. In fact, the RNN(1024U_1L) model outperforms the RNN(1024U_4L) model, indicating that adding more hidden layers does not necessarily improve the results in this case.

The experimental results utilizing RNN+ATT indicate that the RNN+ATT (1024U_2L) achieves the best performance with a CER of 3.72% and a WER of 19.57%. By incorporating an attention mechanism into the RNN-based encoder-decoder model, the performance is improved as it enables better extraction of internal information from the input Mongolian character sequence. Compared with RNN (1024U_1L), RNN+ATT (1024U_2L) has a 2.89% lower WER. The addition of the attention mechanism to the RNN-based encoder-decoder model enables better capture of internal information from the input Mongolian character sequence. This improvement in capturing the internal information leads to enhanced model performance.

Table 3 shows the experimental results of C2T conversion based on the Transformer model. Within the Transformer model, the variable "S"corresponds to the number of layers, while "H"signifies the count of attention heads. The best performing Transformer model is Trans (6S_4H), with a CER of 3.15% and a WER of 16.91%. The model's efficacy is enhanced by increasing the number of attention heads, as shown by the comparison between Trans (6S_4H) and Trans (6S_2H) and Trans (4S_4H) and Trans (4S_2H).

The best results for T2C conversion are achieved by Joint (N=9), RNN (1024U_1L), RNN+ATT (1024U_2L), and Trans (6S_4H) models. Among them, Trans (6S_4H) shows the best performance with a WER of 16.91% and a CER of 3.15%. Compared with Joint (N=9) and RNN models, Trans (6S_4H) had a significantly lower WER. The results suggest that self-attention mechanisms are more effective in capturing the relationship between Mongolian characters and can improve model predictions

5.4. Experimental Results of the combination of rules and Transformer

The evaluation outcomes of the the combination of rules and Transformer method on the "T2C_sentence"dataset are presented below. The test set contains 5,464 OOV words and 93,546 in-vocabulary words. The overall word error rate (WER) on the test set "T2C_sentence"is 18.07%. Among them, the word error rate WER for OOV word conversion based on the Transformer model is 18.13%. Since the text in the test set contains various types of noise, such as typos, abbreviations, and colloquial language, this paper did not use more data augmentation techniques and only performed simple abbreviation restoration operations.

5.5. Experimental Results based on Transformer

The experiments on the "Mon_data"dataset in the Transformer-based C2T conversion showed that both the RNN and Transformer models outperformed the Joint model. Among these models, the RNN model with attention performed better than the RNN model without attention, while the Transformer model exhibited the best performance. However, in the experiments on the "T2C_sentence"dataset, the Transformer model did not achieve optimal results. Further analysis comparing the "Mon_data"and "T2C_sentence"datasets revealed that "T2C_mon"consisted of a significant number of data pairs containing Traditional Mongolian and Cyrillic Mongolian word stems. On the other hand, the "T2C_sentence"test set, the suffixes were not segmented but rather merged with the preceding word, treating them as a single word. As a result, the performance of our trained Transformer model declined on the "T2C_sentence"test set. To address this issue, we expanded the training dataset for the Transformer model. We augmented the original "Mon_data"dataset, which contained 63,668 word pairs, to 108,565 word pairs, creating the augmented dataset called "Mon_data_max". By augmenting the training dataset, we aimed to enhance the performance of the Transformer model on the "T2C_sentence"test set. In the new experiments, we randomly selected 97,709 word pairs from the augmented dataset "Mon_data_max"as the new training set and used 10,856 word pairs as the new test set. On this new test set, the Word Error Rate (WER) of the model was 15.65%. Additionally, the overall WER of the new model on the "T2C_sentence"test set was 16.30%. Specifically, the WER for the out-of-vocabulary word conversion module based on the Transformer model was 16.56%. In summary, by expanding the dataset, the new model achieved good performance on the test set, but there is still room for improvement in handling out-of-vocabulary word conversions.

6. Conclusions

This paper conducted a comparison between RNN and Transformer-based models for character-level C2T conversion tasks. Both models exhibited su perior performance compared to the baseline Joint sequence model, with the Transformer model achieving the highest results. Challenges such as homo phones and inaccurate one-to-many conversions were addressed by combining the Transformer model with Traditional methods. However, the models' per formance was limited by the need for more contextual information. Future research will focus on sentence-level C2T tasks, aiming to leverage contextual information to simplify the conversion process and enhance model performance.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *Computer Science*, 2014.

[2] Feilong Bao, Guanglai Gao, Hongwei Wang, and min Lu. Combining of rules and statistics for cyrillic mongolian to traditional mongolian conversion. 31(3):156, 2017.

[3] Feilong Bao, Guanglai Gao, Xueliang Yan, and Hongxi Wei. Research on conversion approach between traditional mongolian and cyrillic mongolian. *Computer Engineering and Applications*, pages 206–211, 2014.

[4] Maximilian Bisani and Hermann Ney. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5):434–451, 2008, https://doi.org/10.1016/j.specom.2008.01.002

[5] Chaoluomeng. *Modern Mongolian*. Inner Mongolia People's Publishing House, Hohhot, 2009.

[6] Chinggaltai. *A grammar of the Mongolian language*. Inner Mongolia Peoples Publishing House, Hohhot, 1991.

[7] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *Computer Science*, 2014.

[8] She Chuma. *A comparative study of Mongolian and Cyrillic orthography*. Inner Mongolia Education Press, Hohhot, 2010.

[9] Uuganbaatar Dulamragchaa, Sodoo Chadraabai, Byambasuren Ivanov, and Munkhbayar Baatarkhuu. Mongolian language morphology and its database structure. In *2017 International Conference on Green Informatics (ICGI)*, pages 282–285. IEEE, 2017, https://doi.org/10.1109/ICGI.2017.56

[10] Galasamponsige. *Cyrillic Mongolian Learning Book*. Inner Mongolia Education Press, Hohhot, 2006.

[11] Rui Liu, Feilong Bao, Guanglai Gao, Hui Zhang, and Yonghe Wang. Improving mongolian phrase break prediction by using syllable and morphological embeddings with bilstm model. In *Interspeech*, pages 57–61, 2018.

[12] Rui Liu, Berrak Sisman, Feilong Bao, Guanglai Gao, and Haizhou Li. Modeling prosodic phrasing with multi-task learning in tacotron-based tts. *IEEE Signal Processing Letters*, 27:1470–1474, 2020, https://doi.org/10.1109/LSP.2020.3016564

[13] Rui Liu, Berrak Sisman, Feilong Bao, Jichen Yang, Guanglai Gao, and Haizhou Li. Exploiting morphological and phonological features to improve prosodic phrasing for mongolian speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:274–285, 2020, https://doi.org/10.1109/TASLP.2020.3040523

[14] Rui Liu, Berrak Sisman, Jingdong Li, Feilong Bao, Guanglai Gao, and Haizhou Li. Teacher-student training for robust tacotron-based tts. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6274–6278. IEEE, 2020.

[15] Min Lu, Feilong Bao, and Guanglai Gao. Language model for mongolian polyphone proofreading. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 461–471. Springer, 2017. https://doi.org/10.1007/978-3-319-69005-6_38

[16] Martin Popel and Ondřej Bojar. Training tips for the transformer model. *arXiv preprint arXiv:1804.00247*, 2018, https://doi.org/10.2478/pralin-2018-0002

[17] Kanishka Rao, Fuchun Peng, Haşim Sak, and Françoise Beaufays. Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4225–4229. IEEE, 2015.

[18] Bayar Saihan. *Mongolian Dictionary (Cyrillic and Traditional Mongolian Contrastive Dictionary)*. Suoyongbu Printing Press, Hohhot, 2011.

[19] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

[20] Uganbater.D. *Research on Cyrillic and Mongolian script's morphlolgy and conversion system*. PhD thesis, Inner Mongolia University, 2014.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[22] Yonghe Wang, Feilong Bao, Hui Zhang, and Guanglai Gao. Joint alignment learning-attention based model for grapheme-to-phoneme conversion. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7785–7792. IEEE, 2021, https://doi.org/10.1109/ICASSP39728.2021.9413679 PMid:34360079 PMCid:PMC8345426

[23] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017, https://doi.org/10.21437/Interspeech.2017-1452 PMid:28580117 PMCid:PMC5434753

[24] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[25] Zhizhong Zhang. *New Mongolian Chinese Dictionary*. Commercial Press, Beijing, 2011.