Chinese Text De-Colloquialization Technique Based on Back-Translation Strategy and End-to-End Learning

Hongkai Liu 1,2,3,4, Zhonglin Ye 1,2,3,4, Haixing Zhao 1,2,3,4,* and Yanlin Yang 1,2,3,4

1 College of Computer, Qinghai Normal University, Xining 810008, China; hongkai_liu@foxmail.com (H.L.); zhonglin_ye@foxmail.com (Z.Y.); yanlin_yang@foxmail.com (Y.Y.)
2 The State Key Laboratory of Tibetan Intelligent Information Processing and Application, Xining 810008, China
3 Tibetan Information Processing and Machine Translation Key Laboratory of Qinghai Province, Xining 810008, China
4 Key Laboratory of Tibetan Information Processing, Ministry of Education, Xining 810008, China
* Correspondence: h.x.zhao@163.com

Abstract: With the development of the Internet, there has been a significant increase in various types of textual information. However, when people engage in the composition of formal texts, they often incorporate their colloquial habits, which can diminish the professionalism and formality of the text. Existing research on Chinese texts primarily focuses on correcting misspelt characters that are visually or phonetically similar, as well as obvious grammatical errors, such as redundancy, omissions, and incorrect word order. However, there is limited research addressing the correction of text that exhibits colloquial expressions without apparent grammatical errors or misspelt characters. This article proposes a novel technique that utilizes deep learning methods to directly transform colloquial textual expressions into formal written expressions. Firstly, a parallel corpus dataset of written and spoken language is constructed using a back-translation strategy. Then, an end-to-end learning mechanism based on neural machine translation is employed, with colloquial text as the source language and written text as the target language. This allows the model to directly transform the colloquial text into text with a formal style. Finally, an evaluation of the proposed approach is conducted using the bilingual evaluation understudy (BLEU) and manual assessment techniques. The experimental results demonstrate that the technology proposed in this paper performs well in the task of de-colloquialization in Chinese texts. The contribution of this paper lies in proposing an automated method for collecting a substitute for manually annotated parallel corpora of spoken and written language, which significantly saves time and reduces the manual cost of constructing the dataset. Furthermore, the application of end-to-end learning techniques from neural machine translation to the task of de-colloquialization allows the trained model to directly generate written language flexibly based on the input of spoken language. This presents a novel solution for the task of the de-colloquialization of Chinese text.

Keywords: neural machine translation; Chinese texts; de-colloquialism; deep learning; natural language processing

1. Introduction

With the rapid development of the Chinese economy and continuous improvement in China’s comprehensive national strength, Chinese is gaining increasing prominence as an international language. Consequently, increasing amounts of foreign scholars are actively participating in the study of the Chinese language. At present, 10 UN-affiliated agencies have adopted Chinese as the official language, 76 countries have incorporated Chinese into their national education systems, and nearly 200 million people have learned and used Chinese abroad.
The “colloquialization” trend in written English among Chinese students has drawn attention in the field of foreign language teaching. Scholars, such as Wen et al. (2003) [1], Pan (2012) [2], and Han (2013) [3], have investigated this trend using various features. However, overcoming this trend poses a significant challenge in second language education. To address this challenge, a de-colloquialization system can assist Chinese language learners in improving their understanding of Chinese expression, leading to an increasing demand for the use of de-colloquialization in the Chinese language.

De-colloquialization in the Chinese language greatly benefits fields that require extensive text writing in daily life. People may unconsciously include colloquial sentences lacking proper grammatical or spelling errors when writing formal texts due to their habits regarding using colloquial expressions in communication. Furthermore, manual proof-reading is time-consuming and costly. In education, de-colloquialization can improve the language standards and the rigour of textbooks. In the business sector, de-colloquialization can enhance the formality of contracts, showing respect for the cooperating parties. In journalism, de-colloquialization technology can optimize news releases to improve media professionalism and credibility by avoiding colloquial expressions.

The concept of machine translation has been applied to other tasks, as evidenced by the work of Chollampa and Ng (2017), who proposed a grammar correction model based on statistical machine translation, which greatly improved text correction accuracy [4]. One year later, Chollampa and Ng (2018) proposed a correction model using a multi-layered convolutional encode–decode neural network. Due to the architecture design of seq2seq, their model can not only correct grammatical errors but also correct spelling and word priors in the text [5]. This article considers de-colloquialization a translation task, using neural machine translation to “translate” colloquial expressions in text into written expressions. However, in China, there is relatively little research on the de-colloquialization of Chinese text, and no parallel corpus datasets are available to train de-colloquialization tasks. To address this issue, we created a parallel corpus dataset using the back-translation method proposed in 2015 by Facebook [6]. The experiment verifies the effectiveness of the dataset and investigates the performance of two mainstream neural machine translation models: the Sequence to Sequence (Seq2Seq) model (2014) [7] and the Transformer model (2017) [8]. The experiment demonstrates that these models can handle de-colloquialization tasks and produce satisfactory results in human evaluation.

The purpose of this paper is to explore whether the end-to-end translation approach of neural machine translation can be applied to the task of expression style conversion. Furthermore, the goal is to identify the most suitable framework for this task, which can be used for future optimization research. The technique proposed in this paper fills the gap in the study of expression style conversion in natural language processing and lays a foundation for further research in this area.

2. Background
2.1. Colloquialization Research in Chinese

With the increasing number of Chinese language learners, research in the field of language on the colloquialization tendency in written language among learners of Chinese as a second language has moved beyond the level of readers’ intuitive experience. Scholars have started to conduct objective and in-depth research, moving from subjective impressions towards rational analyses.

For example, Ma (2017) investigated the differences in language use between Chinese learners and native speakers in written and spoken contexts by examining the frequency of words [9]. Three features, including modality particles, first- or second-person pronouns, and prepositions, were analyzed to represent language style. Results indicated significant discrepancies in the use of said features between Chinese learners and native speakers in written contexts, with an inclination towards deviating from colloquialism in writing. Luo (2020) summarized the differences between oral and written language at the
level of syllables, vocabulary, and grammar. In addition, compositions by Thai international students were analyzed to verify the tendency for colloquialism in the students’ writing style [10]. Zhang and Song (2011) analyzed and categorized the colloquialization tendencies of Chinese university international students in Shanghai into three main types and eight subtypes, including vocabulary, discourse, and the usage of different levels of Chinese language and characters [11]. Research indicates that the tendency towards colloquialism is a common stylistic feature among second language learners in their writing.

2.2. Neural Machine Translation

Kalchbrenner and Blunsom (2013) proposed a new neural machine translation model using a convolutional neural network to encode the source language and a recurrent neural network (RNN) to decode the target language, marking the birth of neural machine translation [12]. Sutskever et al. (2014) introduced the Seq2Seq method, incorporating long short-term memory (LSTM) structures into neural machine translation to alleviate the problem of gradient explosion. Bahdanau, Cho, and Bengio (2014) first proposed the use of the attention mechanism for neural machine translation, addressing issues of translation under- and over-accuracy through word-to-word alignment [13]. Vaswani et al. (2017) proposed the Transformer method, which uses only multi-headed self-attention and feed-forward neural networks, abandoning RNNs and convolutional neural networks (CNNs). While offering powerful performance, deeper network structures and more parameters require larger parallel corpora for training. Currently, researchers are not only satisfied with the semantic translation of text but also exploring multimodal semantic information, particularly in the domain of image–text matching. Pei et al. (2023) proposed a novel scene graph semantic inference network (SGSIN) for image and text matching. This network effectively learns fine-grained semantic information in both visual and textual modalities, bridging the cross-modal discrepancies [14].

3. Materials and Methods

This paper is inspired by a data-augmentation technique called back-translation, which was used to construct a parallel corpus from a monolingual corpus of formal written language. The constructed parallel corpus was then used to train two state-of-the-art end-to-end translation models, Seq2Seq+attention and Transformer. The trained models were evaluated on a separate validation set to assess their performance. To ensure a fair and objective evaluation, in this paper, we employed two evaluation metrics: BLEU score [15] and manual evaluation. The final experimental results show that the parallel corpus constructed by our proposed method can be used for the task of de-colloquialization and that an effective de-colloquialization model can be trained using the end-to-end machine translation approach. The training process diagram of the proposed method is shown in Figure 1.
Figure 1. Overall flowchart. The collected written text is translated back to colloquial language through back-translation to obtain colloquialized texts that are semantically equivalent. After review, they are compiled into a parallel corpus dataset, from which a portion is extracted as the test set. Then, training is conducted separately on Seq2Seq+attention and Transformer models to obtain two text de-colloquialization models. Subsequently, the test set is predicted using each model, and finally, the model performance is evaluated.

Furthermore, the model trained by our proposed method also possesses some error correction capabilities, such as correcting punctuation and spelling errors. Table 1 illustrates some examples of the model’s error correction abilities.

Table 1. Example of model error correction capability.

<table>
<thead>
<tr>
<th>Language</th>
<th>Original Sentence</th>
<th>De-Colloquialization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>毕业后，我需要曲线救国；这个目标很可能要逐步实施。但是我对整个金融体系还是有信心的“斯考特说：”</td>
<td>“毕业后我需要曲线救国，这个目标可能会逐渐实现，但我对整个金融体系仍然有信心。”斯考特说。</td>
</tr>
<tr>
<td>English</td>
<td>After graduation, I need a curve to save the country; This goal is likely to gradually materialize. But I still have confidence in the entire financial system, ”Scott said:”</td>
<td>After graduation, I need a curve to save the country. This goal may gradually be achieved, but I still have confidence in the entire financial system, ”Scott said.</td>
</tr>
</tbody>
</table>

In the example given above, not only was the colloquial language converted to formal written language, but the model also corrected errors in punctuation. Specifically, the corrected sentence contains a complete set of quotation marks, which were missing in the original sentence. The semicolon after “曲线救国” was changed to a comma, and the period after “逐渐实现” was changed to a comma. Moreover, the colon at the end of the original sentence was changed to a period. In addition to these punctuation corrections, the model also corrected the spelling error “显” to “现”.
3.1. Data Construction Based on Back-Translation

Back-translation is a technique proposed by Sennrich et al. in 2015 to improve the performance of neural network translation models. This technique involves using monolingual corpora to generate synthetic parallel corpora that can be used to enhance the model’s performance. Although formal written Chinese can be easily found in publicly available corpora, it is difficult to find colloquialism that is semantically aligned with the written language. This is why our proposed method uses back-translation to construct a parallel corpus for de-colloquialization. Afterwards, Yu et al. (2018) proposed back-translation as a dedicated data-augmentation technique to optimize the performance of question-answering models [16]. Although the approach to back-translation may differ slightly depending on the specific application, the overall idea remains the same.

The back-translation method for data augmentation translated language 1 into language 2 and then translated language 2 back into language 1. Multiple intermediate languages can also be used. However, to ensure that the semantics of the original language are not altered, it is crucial to have a high-quality translation model that can accurately preserve the semantic content of the text during translation. The effect of back-translation data augmentation using the Google Translate engine is shown in Table 2.

<table>
<thead>
<tr>
<th>Original Sentence</th>
<th>Chinese → Japanese</th>
<th>Japanese → English</th>
<th>English → Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td>一架载有43人的民航飞机周一早间在飞越阿富汗首都喀布尔北部高山时坠毁。当时的天气条件极其恶劣。</td>
<td>アフガニスタンの首都カブールの北にある高山の上空を飛行中、43人が搭乗した民間旅客機が月曜早朝に墜落した。気象条件は非常に悪かった。</td>
<td>Commercial airliner with 43 people on board crashed early Monday morning while flying over high mountains north of the Afghan capital Kabul. Weather conditions were very bad.</td>
<td>周一清晨，一架载有43人的商用客机在飞越阿富汗首都喀布尔以北的高山时坠毁。天气状况非常糟糕。</td>
</tr>
</tbody>
</table>

It can be observed that the back-translated sentences are somewhat different from the original sentences, but their meaning is basically the same.

The back-translation method used in this paper mainly focuses on sentence corpora, that is, how to generate colloquialisms from written language using back-translation. According to the research of Ma, Zhang, Luo, and others, regardless of whether Chinese learners’ mother tongue is English or Thai, they tend to use colloquialisms when writing in Chinese. The nature of neural networks is a biomimetic structure that simulates the neural information transmission inside the human brain. Language is also a kind of information, and some training methods of neural networks are very similar to the way humans learn language. For example, when bidirectional encoder representations from transformers (BERT) train word vectors, they randomly mask a part of the training data, which is very similar to the fill-in-the-blank questions that humans practice when learning English [17]. Therefore, does the language model trained by neural networks also have the characteristic of human colloquialism when generating sentences? We designed experiments to select two mature neural network translation models on the market—A Translator and B Translator. The reason for choosing mature translation models is to ensure that the semantics of the sentences will not change after back-translation and to save the cost of training translation models.

The reason for selecting two different translation models is that when the same translation model is used to train translations in different directions between two languages, the training corpus dataset used is the same, which may cause the back-translated sentences to be similar to the original sentences. To solve this problem, multiple intermediate languages can be selected, as shown in Table 1. Although the original text and the final
translation in Table 1 are different, the degree of colloquialism is not very high because Google Translate’s performance is good enough. Just like in humans, people who are very good at learning Chinese will not speak too colloquially, and those who are worse at Chinese will tend to speak more colloquially. To solve this problem, in this paper, we choose two different translation engines, among which the translation quality of B Translator is slightly lower than that of A Translator, so B Translator is chosen as the translation model for translating back to Chinese through intermediate language translation. In addition, this approach can also save costs because if multiple intermediate languages are selected, each sentence needs to call the translation interface multiple times, which is very complex and troublesome.

When choosing intermediate languages, languages that are farther away from Chinese in terms of grammar structure and word representation should be selected. The reason for this is that the parallel corpus used for training small languages is currently slightly less, which will reduce the performance of the model. The farther away a language is from Chinese, the more features it has, and the harder it is for the neural network to learn, which further reduces the performance of the model. In the case where basic performance has already been maintained, we also do not want the translation model to have too good of a performance so that the back-translated sentences have a greater difference from the original sentences. This is more in line with the requirements of this paper. The back-translation process of this paper is shown in Figure 2.

![Figure 2. Back-translation flowchart. The written language dataset is translated into an intermediate language dataset through A Translator, and then the intermediate language dataset is translated into the corresponding colloquialism dataset through B Translator.](image)

The back-translation effect in this paper is shown in Table 3.

<table>
<thead>
<tr>
<th>Language</th>
<th>Original Sentence</th>
<th>Back-Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>一架载有 43 人的民航飞机周一早间在飞越阿富汗首都喀布尔北部高时坠毁。当时的天气条件极其恶劣。</td>
<td>这架 43 人的民航客机于周一早晨从阿富汗首都喀布尔以北的高山上坠毁。当时的气候条件很恶劣。</td>
</tr>
<tr>
<td>English</td>
<td>A civil aviation plane carrying 43 people crashed on Monday morning while flying over high mountains north of the Afghan capital Kabul. The weather conditions at that time were extremely harsh.</td>
<td>This 43 person civil aircraft crashed from a high mountain north of the Afghan capital Kabul on Monday morning. The climate conditions at that time were very harsh.</td>
</tr>
</tbody>
</table>

In this example, the original sentence exhibits noticeable colloquial features after applying the back-translation method described in this article. Specifically, in the translated sentence, “这架” is used instead of “一架” from the original sentence, and “很” is used in
place of “极其”， which are common colloquial word choices. Additionally, the translated sentence omits modifiers such as “载有” and “飞越”, reflecting the tendency in colloquialism to express ideas more concisely.

3.2. Seq2Seq+Attention

**Seq2Seq**: The Seq2Seq model framework is a sequence training method proposed by Sutskever et al. in 2014, which can map sequences to sequences. This method solves the problem that traditional deep learning networks can only input and output with fixed dimensional encoding. It has been widely used in machine translation, speech recognition, and question-answering systems. Later machine-translation methods were also designed based on this structure.

The basic Seq2Seq model is designed based on RNNs, consisting of two parts: an encoder and a decoder. The initial Seq2Seq model was composed of LSTM cells [18], where one LSTM is utilized to read input sequences time step by time step and acquire a large fixed-dimensional vector representation, and another LSTM is used to extract output sequences from the vector. The second LSTM’s input sequence is conditioned since the decoder will take the previous output as the current input. If the previous output is not sufficiently accurate, it will affect subsequent predictions. Therefore, a technique called teacher forcing is applied during training by using the correct form of each sentence as the input to the decoder, thus enforcing the correct input during training.

**Attention Mechanism**: Although the Seq2Seq model already closely mimics the behaviour and thinking process of human translation, which is to read a sentence in the source language, understand its meaning, and then translate it into the target language in one sentence, there are still two drawbacks to Seq2Seq. Firstly, the encoder struggles to compress a long sentence into a fixed-dimensional vector. Secondly, the decoder forgets information in the output vector generated by the encoder during decoding.

Cho et al. proposed a new approach in 2014 [19] that used the context information vector generated by the encoder as an additional input at each time step during decoding. In 2015, Luong et al. proposed a global attention mechanism for translation [20]. It provides three simple algorithms for calculating attention weights, as shown in Equation (1).

\[
score(h_t, \tilde{h}_s) = \begin{cases} 
  h_t^\top \tilde{h}_s, & \text{dot,} \\
  h_t^\top W_s \tilde{h}_s, & \text{general,} \\
  v_s^\top \tanh(W_s [h_t; \tilde{h}_s]), & \text{concat.}
\end{cases}
\]

In this Equation, \(h_t\) represents the hidden state at the current time step, \(\tilde{h}_s\) represents the hidden state of the source language, and \(W_s\) and \(V_s\) are both parameter matrices that need to be learned.

This paper replaces LSTM with the bidirectional gated recurrent unit (GRU) [21] in the Seq2Seq model. The Seq2Seq+attention de-colloquialism model architecture used in this paper is shown in Figure 3.
Figure 3. Seq2Seq+attention model diagram. The encoder part utilizes a two-layer bidirectional GRU, while the decoder consists of a two-layer unidirectional GRU with the Luong global attention mechanism computed in a general fashion.

3.3. Transformer

Transformer Overall Structure: The Transformer still adopts the encoder–decoder structure, but unlike Seq2Seq, it eliminates the use of convolutional neural networks (CNNs) and RNNs internally and instead employs the self-attention mechanism. In RNNs, the computation of each unit is not parallel, as each unit needs to wait for the computation of other input units requiring their information. Gehring et al. (2017) solved this problem using CNNs, but it requires stacking many layers to capture long-range information [22]. The Transformer can be parallelized directly, and it also follows the encoder–decoder architecture completely.

Encoder: The encoder of the Transformer model consists of two sublayers: the multi-head self-attention layer and the feed-forward neural network layer. The multi-head self-attention layer is composed of multiple attention layers concatenated together, with the attention calculation method in each attention layer using the scaled dot product formula. The formula is in Equations (2) and (3):

\[
Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V'
\]  

(2)

\[
\text{Multi-Head} = [Attention_1; Attention_2; \ldots; Attention_n],
\]  

(3)

In Equation (2), Q, K, and V represent the query matrix, key matrix, and value matrix of the attention layer, respectively, which are obtained by passing the input vector through h different linear layers, representing the feature information of the text in h different dimensions. \(d_k\) represents the dimensionality of vectors.

The formula for the feed-forward neural network layer is as in Equation (4):

\[
FFN = \text{ReLU}(W_1x + b_1)W_2 + b_2,
\]  

(4)

In Equation (4), \(W_1, W_2, b_1, \) and \(b_2\) are the weights and biases of the two linear layers. The process involves expanding the input vector \(x\) to the dimensionality of \(W_1\) and then reducing it back to its original dimensionality using \(W_2\).

After each sublayer, there is a residual connection and normalization to alleviate gradient vanishing caused by the product of many factors during backpropagation. The formula for normalization is as in Equation (5):

\[
SubLayer_{\text{norm}} = \text{LayerNorm}(x + (\text{SubLayer}(x))),
\]  

(5)

In Equation (5), \(x\) represents the vector before processing by the sublayer, \(\text{SubLayer}(x)\) represents the vector after processing by the sublayer, and \(\text{LayerNorm}\) is the normalization function.
**Decoder:** The decoder in this model has two sublayers from the encoder and an additional multi-head attention layer. This enables each decoder position to access all encoder input sequence information, enhancing relevant feature information. To prevent the model from capturing subsequent word information, the input sentence must be masked in the first self-attention sublayer of the decoder. This ensures that the model’s prediction only relies on previous information, maintaining consistency between training and prediction.

4. Results

4.1. Dataset and Preprocessing

The dataset used in this article is a parallel corpus of written and colloquialism, in which the written language was adopted as the target language, and the large-scale Chinese–English translation dataset translation2019zh, publicly available on GitHub via Brightmart, was used for this purpose. This dataset contains 5.2 million Chinese–English sentence pairs, with 5.16 million sentences used as the training set and 39,000 sentences used as the test set. The Chinese sentences in the training set are used as the written language, resulting in 5.16 million Chinese written-language sentences. The colloquialism is generated by back-translation, resulting in 5.16 million colloquialism sentences, which correspond 1-to-1 to the written-language sentences and together form a new parallel corpus dataset with 5.16 million sentence pairs of spoken and written language. The size of the newly created dataset is 970 M. After removing some translation errors through data cleaning, the dataset was reduced to 4.94 million sentence pairs, with a total size of 902 M. The processed dataset was split into training, validation, and testing sets according to a certain proportion, with the size of each data subset shown in Table 4.

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Sentence Pairs</th>
<th>File Size/M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td>4,843,416</td>
<td>883</td>
</tr>
<tr>
<td>Validation set</td>
<td>51,614</td>
<td>9.4</td>
</tr>
<tr>
<td>Test set</td>
<td>49,424</td>
<td>9</td>
</tr>
<tr>
<td>Aggregate</td>
<td>4,944,454</td>
<td>902</td>
</tr>
</tbody>
</table>

Before training on Chinese corpora, the texts need to be tokenized. In this article, the Jieba package’s precise mode was used for Chinese word segmentation. After constructing the dictionary, the word vectors were learned together with the model.

4.2. Experimental Environment

This experiment was conducted on the AutoDL deep learning cloud platform, and the software and hardware environment of the experiment are shown in Table 5.

<table>
<thead>
<tr>
<th>Hardware Environment</th>
<th>Software Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU: 12 vCPU Intel(R) Xeon(R) Platinum 8255C</td>
<td>Operating system: ubuntu20.04</td>
</tr>
<tr>
<td>CPU @ 2.50GHz</td>
<td>PyTorch 1.11.0 + Python 3.8</td>
</tr>
<tr>
<td>Memory: 40 GB</td>
<td>CUDA: 11.3</td>
</tr>
<tr>
<td>Graphics card: RTX 3090ti (24 GB)</td>
<td></td>
</tr>
</tbody>
</table>

4.3. Experimental Parameter Settings

To ensure the maximum performance of the Seq2Seq+attention framework in this article, the parameters were tested, and the batch size was set to 640. The word vector dimensions of the encoder and decoder were set to 300. The sizes of the encoder’s and decoder’s hidden layers were 128 and 356, respectively. The dropout rate of the encoder was
0.3, and the dropout rate of the decoder was 0. The learning rate was set to 0.02, and dynamic decay was applied. During sentence generation, the width of the beam search was set to 15. The Seq2Seq+attention network model was trained for 100 epochs, and the model with the lowest loss value was saved as the final model.

For the Transformer model in this article, both the encoder and decoder had four layers. The batch size was set to 512. The word vector dimensions of the encoder and decoder were both 512. The hidden layer sizes of the encoder and decoder were both 512. The dropout rates for both the encoder and decoder were 0.1. The learning rate was 0.0005. Both the encoder and decoder used eight heads in their multi-head attention mechanisms. The training lasted 60 epochs, and the loss was calculated in the validation set at each training cycle. The model with the lowest validation loss was selected as the final model.

4.4. Evaluating Indicator

In order to ensure the objectivity of the results, this article chose to evaluate the models using two metrics: BLEU score and human evaluation. The essence of the BLEU score is to measure the similarity between two texts, so calculating the written language produced in the test set and comparing it with the original written language can reflect the performance of the model. The formula for BLEU is shown in Equation (6):

$$BLEU = \exp\left(\sum_{n} W_n \log P_n\right)$$

In Equation (6), $P_n$ represents the probability of shared subsequences, specifically, n-grams appearing in the reference sentence, with $n$ up to 4. Thus, only the precision of up to 4 g is measured. $W_n$ represents the weights of each n-gram. $BP$ represents a penalty factor that biases BLEU towards shorter translations. When the predicted sentence is shorter than the reference sentence, it is set to 1. Otherwise, it is dependent on the length ratio between the predicted and reference sentence, as shown in Equation (7):

$$BP = \begin{cases} 1, & \text{if } l_p > l_r, \\ \frac{l_p}{l_r}, & \text{if } l_p < l_r, \\ e^{-\frac{l_r}{l_p}}, & \text{if } l_p = l_r. \end{cases}$$

The final BLEU score is calculated as the average BLEU score of all sentences in the test set. This article calculates BLEU scores for both word-grams and character-grams. However, BLEU is only a substitute for human evaluation, and its disadvantages include the lack of consideration of grammatical accuracy and no consideration of synonyms or similar expressions. Therefore, it is necessary to add a human evaluation as an evaluation criterion for the model.

The manual evaluation method used in this article is in the form of a questionnaire survey. Each questionnaire was composed of 500 non-colloquial sentence pairs randomly selected with a replacement from the test set. They were distributed to 20 graduate students at our university to evaluate the effectiveness of de-colloquialization for each sentence on a ten-point scale. The scores were averaged for each questionnaire, and the average scores across all questionnaires determined the final de-colloquialization score for the model.

4.5. Result and Analysis

This article employs two outstanding models in the field of machine translation for training to investigate whether the principles of machine translation can be applied to the task of de-colloquialization. It is observed that the Seq2Seq+attention model failed to fit the dataset for de-colloquialization well. Even after 100 epochs of training, the loss value remained as high as 5.77. To verify whether this phenomenon is attributed to the large size of the dataset, this article resorts to 1 million parallel corpora to further train the
model. However, even after 100 epochs, the loss value reached 5.15, as shown in Figure 4, indicating a limited improvement in the loss value.

Figure 4. Seq2Seq+attention training process. Schemes follow the same formatting. It can be observed that after 20 epochs, the loss value starts to descend at a very slow pace. Even though the model is able to continue converging after 100 epochs, its convergence efficiency is considerably reduced.

In contrast, the Transformer model converges to remarkable effectiveness within only 60 epochs, as shown in Figure 5.

Figure 5. Transformer training process. The process of how the loss values of the Transformer model on both the training and validation sets decrease is shown in (a), while the perplexity score of the Transformer model is shown in (b). After being trained for 60 epochs, the model’s loss value is as low as 1.49, and the validation loss value also decreases to 1.42. Furthermore, the perplexity score of the model experiences a significant drop.

The comparison of experimental results is presented in Table 6.

Table 6. Comparison of experimental results.

<table>
<thead>
<tr>
<th>Model</th>
<th>Data</th>
<th>Loss</th>
<th>BLEU (Word)</th>
<th>BLEU (Character)</th>
<th>Manual</th>
<th>Time/h</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial BLEU</td>
<td>4.9 W</td>
<td>/</td>
<td>24.9</td>
<td>34.1</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Seq2Seq</td>
<td>500 W</td>
<td>5.88954</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>70.21</td>
</tr>
<tr>
<td>Seq2Seq+attention</td>
<td>100 W</td>
<td>5.15372</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>49.11</td>
</tr>
<tr>
<td>Seq2Seq+attention</td>
<td>500 W</td>
<td>5.77583</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>209.3</td>
</tr>
<tr>
<td>Transformer(base)</td>
<td>500 W</td>
<td>1.62311</td>
<td>26.2</td>
<td>35.3</td>
<td>6.8</td>
<td>30.3</td>
</tr>
<tr>
<td>Transformer</td>
<td>500 W</td>
<td>1.49185</td>
<td>27.3</td>
<td>36.4</td>
<td>7.36</td>
<td>36.6</td>
</tr>
</tbody>
</table>
Table 6 displays detailed metrics for the training of both the Seq2Seq+attention and Transformer models. These metrics include the amount of data used for training, the final converged loss value, the character-level and word-level BLEU scores (a character-matching algorithm compared to the reference answers), the human evaluation score, and the training time. The initial line of data aims to showcase the computation of BLEU scores for a test set comprising 49,000 paired spoken and written texts. These scores are considered fundamental indicators capable of reflecting the similarity between spoken and written texts. They are utilized for comparison with the BLEU scores of sentences that have undergone model transformation. It is observed that, for sentences transformed by the Transformer model, whether at the character-level BLEU or word-level BLEU, all exceed the baseline indicator. This indicates that the sentences transformed by the model are more closely aligned with written language. Seq2Seq is a pure translation model without an attention mechanism. In the case of the Transformer (base) model, it has four attention heads in the multi-head attention mechanism and a hidden layer dimension of 256. On the other hand, the Transformer model has eight attention heads in the multi-head attention mechanism and a hidden layer dimension of 512.

As the Seq2Seq+attention model did not converge to an ideal state, it was excluded from the model evaluation. It is notable that the Transformer model outperforms the Seq2Seq+attention model in terms of convergence speed and training duration.

Regarding the task of de-colloquialization, since both the target and source languages are Chinese, the difference between the grammar and semantics of the two languages is relatively small. At this point, the multi-head attention mechanism can effectively consolidate features at various dimensions, while Seq2Seq+attention may struggle to learn useful features in a single dimension and could forget some relevant features. Therefore, the Seq2Seq+attention model may not perform well in converging within linguistic datasets with a smaller degree of language distance.

The reason why the calculated BLEU score is very high is that the source language and the target language are already very similar. For delignifying Chinese, although there are a lot of changes in word order and synonym replacements, they did not have a significant impact on the calculation of BLEU.

The Transformer model achieved good ratings in the manual evaluation score indicator. De-colloquialization is a relatively subjective task, and each individual may have different perceptions of whether a sentence is colloquial or not, resulting in varying scores. The de-colloquialization effects of different types of colloquial sentences in the Transformer model are shown in Tables 7–11. In order to facilitate readers’ comprehension, an English translation of the Chinese text has been provided. It is important to note that due to the differences between English and Chinese in terms of colloquial expressions, the translation of colloquial expressions may not be entirely accurate in order to convey the original meaning of the Chinese text. The English translation is offered solely for the purpose of assisting readers and should be used as a reference only.

Table 7. Example of de-colloquialization of noun phrases.

<table>
<thead>
<tr>
<th>Language</th>
<th>Original Sentence</th>
<th>De-Colloquialization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>父母是我第一任老师，谢谢爸妈。 (爸妈)</td>
<td>父母是我的第一任老师，感谢父母。</td>
</tr>
<tr>
<td>English</td>
<td>My parents were my first teachers.</td>
<td>My parents were my first teachers,</td>
</tr>
<tr>
<td></td>
<td>Thanks, Mom and Dad. (Mom and Dad)</td>
<td>and I express my gratitude to my parents.</td>
</tr>
</tbody>
</table>

This example demonstrates the model’s capability in de-colloquializing nouns. In the given sentence, “谢谢” and “爸妈” are colloquial terms in Chinese, where “谢谢” is used as a colloquial verb and “爸妈” is a colloquial noun. After applying the de-colloquialization model, “谢谢” is replaced with “感谢”, and “爸妈” is replaced with “父母”.
### Table 8. Example of de-colloquialization of verb phrases.

<table>
<thead>
<tr>
<th>Language</th>
<th>Original Sentence</th>
<th>De-Colloquialization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>干什么事情都先考虑很长时间才决定。（干）</td>
<td>做什么事情都先考虑很长时间才决定。</td>
</tr>
<tr>
<td>English</td>
<td>Taking a long time to think before doin’ anything. (doin’)</td>
<td>Taking a long time to consider before doing anything.</td>
</tr>
</tbody>
</table>

This example showcases the model’s ability to de-colloquialize verbs. In the given sentence, “干” is a colloquial verb commonly used in Chinese. After applying the de-colloquialization model, “干” is replaced with “做”, which is a more formal equivalent.

### Table 9. Example of de-colloquialization of conjunctions.

<table>
<thead>
<tr>
<th>Language</th>
<th>Original Sentence</th>
<th>De-Colloquialization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>《我对男女分班的看法》不光在中国有人谈论，连在泰国近期也谈到。 (不光…连…)</td>
<td>《我对男女分班的看法》不仅在中国有人谈论，而且在泰国近期也谈到了。</td>
</tr>
<tr>
<td>English</td>
<td>“My views on gender-segregated classes” is not just a topic of discussion in China, but even in recent times, it has been talked about in Thailand.</td>
<td>Not only have “My views on gender-segregated classes” been discussed in China, but moreover, they have also been talked about in Thailand recently.</td>
</tr>
</tbody>
</table>

This example demonstrates the model’s ability to de-colloquialize conjunctions. In the given sentence, “不光 … 连 …” is a colloquial expression commonly used in Chinese to indicate a progressive relationship. After applying the de-colloquialization model, “不光 … 连 …” is replaced with “不仅 …而且 …”, which is a more formal expression to convey a progressive relationship.

### Table 10. Example of de-colloquialization of modal particles.

<table>
<thead>
<tr>
<th>Language</th>
<th>Original Sentence</th>
<th>De-Colloquialization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>我不太习惯，可是现在呢，好像开始习惯了。</td>
<td>我不太习惯，但现在，好像已经开始习惯了。</td>
</tr>
<tr>
<td>English</td>
<td>I’m not really used to it, but like, now, it’s like I’m starting to get the hang of it, you know?</td>
<td>I’m not really used to it, but now, it seems like I’ve already started getting used to it.</td>
</tr>
</tbody>
</table>

This example showcases the model’s ability to de-colloquialize modal particles, which are only found in colloquialism. In the given sentence, both “呢” and “吧” are modal particles in Chinese used to emphasize the emotional tone of expression. After applying the de-colloquialization model, the model not only removes the modal particles “呢” and “吧”, but also modifies the sentence to retain its original meaning. For example, “可是” is changed to “但” and “好像” is followed by “已经”.

### Table 11. Example of de-colloquialization of grammar.

<table>
<thead>
<tr>
<th>Language</th>
<th>Original Sentence</th>
<th>De-Colloquialization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>这篇文章不仅证实了关于两个距离的不动的理由，这一结果推广到了 $K_2$ 个距离。</td>
<td>本文证明了两个距离的不动点的存在性定理，并将这一结果推广到 $k_2$ 距离。</td>
</tr>
</tbody>
</table>


This article demonstrates the existence theorem for stationary points of the two distances and generalizes this result to the $k_2$ distance.

<table>
<thead>
<tr>
<th>Original</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>English</strong></td>
<td>This article right here not only confirms the reasons why those two distances don’t change, but it also applies that finding to K-2 distances. This article demonstrates the existence theorem for stationary points of the two distances and generalizes this result to the $k_2$ distance.</td>
</tr>
</tbody>
</table>

This example highlights the model’s ability to de-colloquialize in terms of grammar. In the original sentence, the conjunction “不仅” lacks a fixed collocation, and the phrase “两个距离的不动的理由” is unclear and lacks coherence. After applying the de-colloquialization model, the model completely restructures and summarizes the content, resulting in “两个距离的不动点的存在性定理”, which means “the existence theorem of two fixed points of distances”. The model removes the conjunction “不仅” and uses “并” to indicate a progressive relationship.

5. Discussion

This article utilizes the back-translation approach to construct a parallel corpus dataset of written language and colloquialisms for training purposes. Regarding the question of why written language, after undergoing back-translation, can produce sentences with colloquial characteristics, the following discussion is provided:

- In this article, the same translation model is not used for back-translation. Instead, two translation models, A and B, are selected for the back-translation process, with translation model B being inferior to translation model A. This is performed to simulate the scenario where humans perform translations. The two language models represent individuals with different native languages and translation proficiency levels. When written language is translated into an intermediary language using translation model A, we aim for the intermediary language to better preserve the original intent of the written language. Therefore, translation model A, known for its better translation performance, is used for this step. It simulates a translator with a higher proficiency level who tends to use advanced vocabulary and expressions. Next, the intermediary language is translated back into the source language using the second translation model, B. In this step, we intentionally introduce noise to create colloquial features. Therefore, translation model B, known for its poorer translation performance, is selected. This step simulates a translator with a lower proficiency level who avoids using advanced vocabulary and expressions but aims for a more colloquial and informal style. Consequently, in this article, we employ this method to simulate the translation process performed by humans and generate noise that introduces colloquial features. Since both translation models are well trained, there is no concern about significant changes in the semantic meaning of the translated sentences.

- When selecting an intermediary language, it is advisable to choose a low-resource language that is significantly different from the source language in terms of grammar and lexical representation. This is because the greater the distance between the intermediary language and the source language, the more challenging the translation becomes, resulting in a higher level of noise that introduces colloquial features. Additionally, selecting a low-resource language is preferred because there is currently limited training data available for such languages. As a result, the performance of the translation model is further weakened, allowing for the generation of more noise during the back-translation process.

In terms of model training, this paper selects two different machine translation models, Seq2Seq+attention and Transformer, and compares their fitting performance to determine the optimal model for subsequent work. It was found that when training on a dataset of 5 million parallel sentence pairs, Seq2Seq+attention was unable to fit the data, while Transformer achieved good results after only 60 batches. To investigate whether the inability to fit the data was due to the large dataset, Seq2Seq+attention was trained on a smaller dataset of 1 million parallel sentence pairs, but the results still showed a lack of
fitting. This may be because Seq2Seq+attention was designed for different source and target languages with significant feature differences, enabling it to effectively capture those features. However, in the task of text de-colloquialization, both the source and target languages are the same, making it difficult for Seq2Seq+attention to effectively capture colloquial features. The disparity may be attributed to the introduction of the multi-head attention mechanism in the Transformer model, which enables effective feature extraction across different linear spaces and various dimensions.

Due to limitations in devices and datasets, the full potential of the models in this experiment was not fully realized. Future work will involve expanding the database by introducing sentence noise such as word order errors, grammar errors, spelling mistakes, and missing words. This will enhance the model’s proofreading ability. Additionally, it will be necessary to develop a metric that effectively measures the degree of colloquialization in a sentence. The ultimate goal is to train a multifunctional language model with text proofreading and de-colloquialization capabilities. Furthermore, future research will consider fine-tuning the models using pre-trained language models to enhance their understanding and further improve performance.

6. Conclusions

In this work, a new method for text de-colloquialism is proposed, and the feasibility of this method is demonstrated. From the construction of the dataset to the training of the model, a back-translation approach is utilized to introduce colloquialism characteristics as noise into written language, thereby obtaining parallel corpora and significantly reducing the cost of manual data collection. By employing the end-to-end translation concept in machine translation, where colloquialism is treated as the source language and written language as the target language, the model is directly trained to perform the conversion between the two languages. As a result, an effective text de-colloquialism model is obtained.

For the task of text de-colloquialism, two of the best translation models were selected for experimentation. The Seq2Seq+attention model did not fit the dataset well, while the Transformer model achieved good results in a shorter time. This demonstrates that the Transformer model not only possesses powerful learning capabilities, but its parallel structure also accelerates the training process. The text de-colloquialism model trained with Transformer achieved a character-level BLEU score of 36.4 and a word-level BLEU score of 27.3, indicating that its de-colloquialism performance is approaching the standard reference. Furthermore, in a human evaluation on a scale of 1 to 10, the model scored 7.6 points.

Text de-colloquialism is a highly meaningful task, despite the current lack of substantial research in this area. Building upon the foundation established in this paper, we will continue to conduct further in-depth research. Our next step will involve designing an evaluation metric that can quantitatively measure the degree of colloquialism in sentences. This metric will provide an intuitive reflection of the level of colloquialism in sentences, eliminating the need for subjective judgments by humans. We will also continue to optimize the content of the dataset and improve the model to achieve better performance in text de-colloquialism tasks.

Author Contributions: Conceptualization, H.L.; methodology, H.L.; software, H.L.; validation, H.L., H.Z., Z.Y., and Y.Y.; formal analysis, H.L.; investigation, H.L.; resources, H.Z. and Z.Y.; data curation, H.L. and Y.Y.; writing—original draft preparation, H.L.; writing—review and editing, H.Z., Z.Y., and Y.Y.; visualization, H.L.; supervision, H.Z. and Z.Y.; project administration, H.Z. and Z.Y.; funding acquisition, H.Z. and Z.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China (No.2020YFC1523300), Innovation Platform Construction Project of Qinghai Province (2022-ZJ-T02).
Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to ongoing follow-up research.

Conflicts of Interest: The authors declare no conflict of interest.

References


Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.