Reducing Multilingual Context Confusion for End-to-end Code-switching Automatic Speech Recognition

Shuai Zhang\textsuperscript{1,2}, Jiangyan Yi\textsuperscript{2}, Zhengkun Tian\textsuperscript{1,2}, Jianhua Tao\textsuperscript{1,2,3}, Yu Ting Yeung\textsuperscript{4}, Liqun Deng\textsuperscript{4}

\textsuperscript{1}School of Artificial Intelligence, University of Chinese Academy of Sciences, China
\textsuperscript{2}NLPR, Institute of Automation, Chinese Academy of Sciences, China
\textsuperscript{3}CAS Center for Excellence in Brain Science and Intelligence Technology, China
\textsuperscript{4}Huawei Noah’s Ark Lab, Shenzhen, China

\{shuai.zhang, jiangyan.yi, zhengkun.tian, jhtao\}@nlpr.ia.ac.cn, \{yeung.yu.ting, dengliqun.deng\}@huawei.com

Abstract

Code-switching deals with alternative languages in communication process. Training end-to-end (E2E) automatic speech recognition (ASR) systems for code-switching is especially challenging as code-switching training data are always insufficient to combat the increased multilingual context confusion due to the presence of more than one language. We propose a language-related attention mechanism to reduce multilingual context confusion for the E2E code-switching ASR model based on the Equivalence Constraint (EC) Theory. The linguistic theory requires that any monolingual fragment that occurs in the code-switching sentence must occur in one of the monolingual sentences. The theory establishes a bridge between monolingual data and code-switching data. We leverage this linguistic theory to design the code-switching E2E ASR model. The proposed model efficiently transfers language knowledge from rich monolingual data to improve the performance of the code-switching ASR model. We evaluate our model on ASRU 2019 Mandarin-English code-switching challenge dataset. Compared to the baseline model, our proposed model achieves a 17.12\% relative error reduction.

Index Terms: Automatic Speech Recognition, Code-Switching, End-to-End, Multilingual Context Confusion

1. Introduction

Code-switching refers to the use of multiple languages in communication, which is a common language phenomenon especially for bilingual speakers [1]. E2E ASR models achieve excellent performance on monolingual speech recognition with joint optimization of acoustic, pronunciation, and language models [2, 3, 4, 5, 6, 7]. Training an E2E code-switching ASR system is still challenging as it heavily relies on audio-text data pairs [8, 9, 10, 11]. The presence of more than one language in an utterance increases the chance of multilingual context confusion.

To alleviate the problem of multilingual context confusion, a natural idea is to use a large amount of monolingual audio-text data to train the code-switching ASR model. Although large datasets significantly improve the performance of monolingual ASR tasks with E2E models [3, 4, 5], it is still insufficient to handle code-switching problem with conventional E2E models. A possible reason is that multilingual context information is not aligned from monolingual data of different languages during model training. Code-switching text data augmentation technology tries to alleviate the multilingual context confusion problem. The method often applies a translation system to artificially generate code-switching text from monolingual text according to various rules [12, 13]. The generated text data increase the richness of multilingual context, and improve the system’s ability to model knowledge of different languages. To generate more reasonable code-switching text, various linguistic theories are proposed to constrain the generation strategy [14, 15, 16]. There are many attempts to explain the grammatical constraints on code-switching. The three most widely accepted theories including the Embedded Matrix (EM) [17], the Equivalence Constraint (EC) [18] and the Functional Head Constraint (FHC) theories [19]. Although applying linguistic theories to text data generation improves the performance of code-switching ASR, text data only assist E2E ASR models indirectly in the form of language models [20, 21, 22]. Moreover, text data generation requires extra computational cost and increases the complexity of ASR training pipeline. Therefore, there is great benefit if we can directly apply linguistic theories to an E2E code-switching ASR model.

We propose a language-related attention mechanism to reduce multilingual context confusion for E2E code-switching ASR model based on the EC linguistic theory. Our method is based on two viewpoints. First, language switching in code-switching is extremely random. There are almost infinite switching combinations. Trying to cover all combinations through text generation is unrealistic. Second, the EC linguistic theory states that any monolingual fragment that occurs in a code-switching sentence must occur in one of the monolingual sentences. This linguistic theory establishes a bridge between monolingual data and code-switching data. On this basis, we propose a method to reduce multilingual contextual confusion that with monolingual data. Specifically, our method is based on transformer structure, which has achieved outstanding performance in the field of ASR. This method uses self-attention mechanism to model language context information, as in the field of natural language processing. Due to the randomness of language switching, attention calculations between different languages are difficult and aggravate the confusion between contexts. Therefore, we adjust the attention calculation method between different languages to make the model more emphasis on the context of the same language. This strategy can reduce the complexity of multilingual context and utilize rich monolingual data more efficiently. According to the EC theory, our method theoretically covers most of language contexts.

Here are our main contributions in this work. First, we propose a language-related attention mechanism to reduce multilingual context confusion for E2E code-switching ASR model based on the EC linguistic theory. Second, we propose three
The paper is organized as follows. In Section 2, we briefly review the structure of Speech-Transformer ASR model. In Section 3, we introduce the language-related attention strategies of our method in details. In Section 4, we describe our experimental design and discuss the experimental results. Finally, we conclude this paper in Section 5.

2. Speech-Transformer

In order to describe our method in a clear manner, we first review the structure of Speech-Transformer model [5]. Speech-Transformer is a successful ASR transformer-based model [24], which is composed of an encoder and a decoder.

The encoder encodes and transforms acoustic features to high-level representation. First, we apply a convolutional neural network (CNN) module to down-sample the acoustic feature sequence and obtain an initial hidden representation. Then we apply several layers of encoder blocks of the same structure to further encode the down-sampled acoustic representation. Each encoder block consists of two sub-blocks: multi-head self-attention and position-wise feed-forward layer. Layer normalization and residual connections are applied between sub-blocks to stabilize training process and improve performance.

The decoder receives the output of the encoder and performs a loss calculation by matching the acoustic representation with target text. First, we apply word embedding layer to convert discrete modeling units into vector representations. Then we apply a stack of decoder blocks to encode the text context and interact with the encoded acoustic representation subsequently. The decoder block consists of three parts: masked multi-head self-attention, multi-head cross-attention and position-wise feed-forward layer. The masked multi-head self-attention models language context information, which is the focus of this paper. The queries, keys, and values are the word embedding representation. The multi-head cross-attention completes the information interaction between the decoder and the encoder through the attention mechanism. The rest of the decoder structure is similar to the encoder.

3. Language-Related Attention Mechanism

According to the EC linguistic theory, any monolingual fragment that occurs in a code-switching sentence must occur in one of monolingual sentences. This linguistic theory establishes a bridge between monolingual data and code-switching data. General E2E models cannot effectively model such linguistic constraints. Specifically, for transformer model, the multi-head self-attention module of the decoder learns language context information. The module uses the same attention calculation method for different language modeling units. Different languages are treated as the same language. Multilingual contextual information between languages cannot be obtained from monolingual data. Monolingual data sometimes even reduce performance of the code-switching ASR model. Therefore, we modify the self-attention of the decoder to strengthen contextual relationship of the same language and reduce contextual relationship between different languages. Multilingual contextual knowledge can be learned more effectively from rich mono-
lingual data. We have implemented three specific language-
related attention schemes. The corresponding model structures are shown in Fig. 1.

As shown in Fig. 1(a), we re-weight the self-attention scores according to the language. Specifically, when the at-
tention scores of a modeling unit are calculated with respect
to other units in the sentence, the attention scores belonging to
the same language are increased while the attention scores of
different language modeling units are reduced. In this way, a
stronger connection is established between the modeling units
of the same language. The mutual influence between differ-
ent languages is suppressed. Therefore, language context repre-
sentation learned from code-switching data is more compatible
with monolingual data. The model can extract context informa-
tion from monolingual data more effectively.

The calculation of the attention is expressed as,

\[
\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}V\right)
\]  

where \(Q, K, V\) denote the query, key, and value respectively,
\(d_k\) is the dimension of the key. For the self-attention, the query,
key, and value are the target text embedding. Our method is
formally expressed as

\[
\text{OurAttention}(Q, K, V, W) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}WV\right)\]

where \(W\) denotes the re-weighting matrix, which is obtained
according to the language switching distribution in the sam-
ple. The matrix completes the process of calculating language-
related attention by adjusting the word embedding matrix or the
attention score matrix.

In addition to adjusting the attention scores, we also imple-
ment two other different attention methods. As shown in Fig.
1(b), we separate the code-switching embedding representation
by languages and get the embedding sequences of the two lan-
guages. Then these two sequences are treated as monolingual
cases for attention calculation. In Fig. 1(b), the two sequences
share the same attention calculation parameters. After obtaining
the corresponding context representations, the representations
of the two languages are merged to obtain a complete code-
switching text representation. This approach is able to achieve
the goal of strengthening the connection with the same language
while reducing the interference of different languages.

The third method described in Fig. 1(c) is similar to the
one in Fig. 1(b), except that the two language embedding se-
quences have their own independent self-attention parameters.
The design of this model structure minimizes mutual interfer-
ence between the two languages. This is a more thorough way
of calculating language-related attention and theoretically can
make use of monolingual data more effectively. After complet-
ing the self-attention calculation, the context fusion of the two
languages is the same as Fig. 1(b). Note that the subsequent
cross attention calculation is completely consistent with the or-
dinary Speech-Transformer model.

In the training stage, three implementation schemes men-
tioned above require the participation of a re-weighting matrix,
which is generated according to the target text during the data
reading phase. In the test stage, the matrix is generated and
expanded in real time with the auto-regressive decoding pro-
cess. The method does not perform language identification in
advance, but only on the basis of the already decoded results.

4. Experiments

4.1. Datasets

We perform experiments with ASRU 2019 Mandarin-English
code-switching challenge dataset. The corpus consists of about
200 hours code-switching training data and 500 hours monolin-
gual Mandarin training data [23]. The audio data are collected
by mobile phones in quiet environments with 16kHz sampling
rate. To facilitate experimental design, we further include the
460-hour subset of Librispeech English dataset [25] into train-
ing set. The development set and the test set each consists of 20
hour code-switching data.

4.2. Experiment Setups

We apply 40-dimensional filter-banks with frame length of 25
ms and frame shift of 10 ms as acoustic features. The modeling
units for Chinese and English are characters and word pieces re-
spectively. We choose 1.0k English word pieces as the English
modeling units. For Chinese, we reserve characters with more
than 10 occurrences in the training set as modeling units. There
are about 3,1k Chinese characters in total. We apply mix error
rate (MER) as the evaluation metric. It counts word error for
English and character error for Chinese. In monolingual cases,
we compute word error rate (WER) for English and character
error rate (CER) for Chinese. The metric is widely adopted to
evaluate the Mandarin-English code-switching ASR system.

We implement all the models with Speech-Transformer ar-
chitecture. We down-sample the acoustic features with two 3×3
2D-CNN layers with stride 2. The dimension of the sub-
sequent linear layer is 512. We apply relative position encoding
to model position information. We use a convolution-augmented
transformer (conformer) [7] as the acoustic encoder with di-

dimension 512. The number of attention heads is 4. The size
of the convolution kernel is 32. For decoder, the attention di-

mension is 512. The number of the head is 4. The dimension of
position-wise feed-forward networks is 1024. There are 12 and
6 blocks for encoder and decoder respectively. In our method,
we need to separate two monolingual embeddings from mixed
embeddings. Specifically, by weighting the embedding of one
language, we obtain the embedding representation of another
language. The weighting parameter is set to 0.1. To further
improve the performance of the model, the weighted sum of
the connectionist temporal classification (CTC) loss and cross-
entropy loss is used as the final loss function. The weighting
parameter of CTC loss is set to 0.2. The weight parameter of
the CTC score during decoding is set to 0.3.

We apply uniform label smoothing [26] to avoid over-
fitting. The parameter is set to 0.1. We also apply SpecAug-
ment with frequency masking (F=30, mF=2) and time masking
(T=40, mT=2) to improve robustness of the models [27]. We
apply residual connections between sub-blocks for stable train-
ing [28]. We set the parameter of residual dropout to 0.1 [29].
The model is trained using 2 NVIDIA V100 GPUs with the opti-
mization strategy of \(\beta_1 = 0.9, \beta_2 = 0.998, \epsilon = 1e^{-8}\) [5].
The batch-size is set to 32. The learning rate is set by a warm-
up strategy [24]. We apply model averaging to improve model
performance. During beam-search decoding, the beam-width is
set to 10.

4.3. Performance with Code-switching Data

We design a comparative experiment between the proposed
method and the baseline model based on the code-switching
data. The experimental results are shown in Table 1. The at-

3896
Table 1: MER/CER/WER (%) of our methods and baseline model with only code-switching data. MER refers to whole utterances. CER refers to Chinese part and WER refers to the English part. ‘score re-weighted’, ‘attention shared’ and ‘attention not shared’ respectively correspond to the aforementioned three self-attention adjustment schemes.

<table>
<thead>
<tr>
<th>model</th>
<th>Data</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MER</td>
<td>CER</td>
<td>WER</td>
</tr>
<tr>
<td>baseline</td>
<td>12.08</td>
<td>9.85</td>
<td>30.04</td>
</tr>
<tr>
<td>score re-weighted</td>
<td>12.15</td>
<td>9.93</td>
<td>30.03</td>
</tr>
<tr>
<td>attention shared</td>
<td>11.40</td>
<td>9.28</td>
<td>28.54</td>
</tr>
<tr>
<td>attention not shared</td>
<td><strong>11.09</strong></td>
<td>9.00</td>
<td>28.01</td>
</tr>
</tbody>
</table>

Table 2: MER/CER/WER (%) of transformer baseline with extra monolingual data. ‘200h CS’, ‘500h CH’, ‘460h EN’ respectively refer to code-switching data, Chinese data, English data and ‘All’ corresponds to the above three datasets.

<table>
<thead>
<tr>
<th>model</th>
<th>Data</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MER</td>
<td>CER</td>
<td>WER</td>
</tr>
<tr>
<td>baseline</td>
<td>12.08</td>
<td>9.85</td>
<td>30.04</td>
</tr>
<tr>
<td>+ 500h CH</td>
<td>11.48</td>
<td>9.00</td>
<td>31.43</td>
</tr>
<tr>
<td>+ 460h EN</td>
<td>11.99</td>
<td>10.01</td>
<td>28.03</td>
</tr>
<tr>
<td>All</td>
<td><strong>11.19</strong></td>
<td>8.99</td>
<td>28.87</td>
</tr>
</tbody>
</table>

Table 3: MER/CER/WER (%) of self-attention sharing strategy with extra monolingual data.

<table>
<thead>
<tr>
<th>model</th>
<th>Data</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MER</td>
<td>CER</td>
<td>WER</td>
</tr>
<tr>
<td>200h CS</td>
<td>11.40</td>
<td>9.28</td>
<td>28.54</td>
</tr>
<tr>
<td>attention + 500h CH</td>
<td>9.77</td>
<td>7.42</td>
<td>28.63</td>
</tr>
<tr>
<td>shared + 460h EN</td>
<td>11.20</td>
<td>9.35</td>
<td>26.08</td>
</tr>
<tr>
<td>All</td>
<td><strong>9.63</strong></td>
<td>7.51</td>
<td>26.74</td>
</tr>
</tbody>
</table>

Table 4: MER/CER/WER (%) of self-attention independence strategy with extra monolingual data.

<table>
<thead>
<tr>
<th>model</th>
<th>Data</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MER</td>
<td>CER</td>
<td>WER</td>
</tr>
<tr>
<td>200h CS</td>
<td>11.09</td>
<td>9.00</td>
<td>28.01</td>
</tr>
<tr>
<td>attention + 500h CH</td>
<td>9.40</td>
<td>7.03</td>
<td>28.54</td>
</tr>
<tr>
<td>not shared + 460h EN</td>
<td>11.08</td>
<td>9.31</td>
<td>25.31</td>
</tr>
<tr>
<td>All</td>
<td><strong>9.34</strong></td>
<td>7.25</td>
<td>26.18</td>
</tr>
</tbody>
</table>

tention scores re-weighting strategy slightly degrades the performance of the model. The performance degradation may be due to the confusion of attention caused by the artificial post-adjustment of the attention scores. The other two attention adjustment strategies improve the performance of the model to a certain extent. Due to the small amount of code-switching data, this experimental result is unable to fully reflect the superiority of our method. The advantage of our method is the ability of using monolingual data.

4.4. Performance with Extra Monolingual Data

To demonstrate the superiority of the proposed method in utilizing monolingual data, we perform experimental comparisons between the baseline model and our methods with additional monolingual data. The experimental results of the baseline model with extra monolingual data are listed in Table 2. The results show that when we add monolingual data, recognition error rate of the corresponding language is reduced. However, this action leads to negative impact on other language. A possible reason is mutual interference between the two language data. Adding monolingual data does not always improve the MER of code-switching ASR task for the ordinary Speech-Transformer model. The baseline model is relatively inefficient in improving code-switching ASR performance using extra monolingual data. Finally, the performance of the baseline model with all training data achieves 7.01% relative MER reduction compared with the results of only code-switching data.

Then we conduct experiments on our proposed methods with the same data design scheme. Since the experimental results of the self-attention scores re-weighting strategy are not satisfactory and the limited paper space, we only show the experimental results of the other two methods. Table 3 shows the experimental results of the self-attention sharing strategy. This method achieves greater improvement with extra monolingual data than the baseline model. Compared with the baseline model, the degradation of extra monolingual data to the performance of the competing language is greatly reduced. The results support our claims that our method can reduce context interference between the two languages and improve performance of code-switching ASR model. The results also demonstrate that our method is more effective in using monolingual data. Finally, the performance of the proposed model with all training data achieves 16.28% relative MER reduction compared with the results trained with only code-switching data.

Table 4 shows the experimental results of the self-attention independence strategy. The experimental results are consistent to Table 3. The strategy with independent self-attention for each language is more effective and achieves the best recognition performance. Since the two languages carry out context modeling independently, mutual interference further reduces. Overall, the proposed model achieves 17.91% relative reduction in MER compared with the results of only code-switching data. Compared with the baseline model, the proposed method achieves 17.12% relative reduction in MER. The above experimental results suggest that our method is an effective solution for code-switching ASR tasks.

5. Conclusion

In this paper, we propose a language-related attention mechanism to reduce multilingual context confusion for the E2E code-switching ASR model based on the EC linguistic theory. These methods can reduce mismatch between code-switching data and monolingual data in modeling language context. Experimental results suggest that the proposed methods utilize monolingual data effectively to improve the performance of code-switching ASR task, compared to the baseline model.

6. Acknowledgment

This work is supported by the Key Research Project of China (No.2019KD0AD01), the National Natural Science Foundation of China (NSFC) (No.61901473, No.62101553, No.61831022). This research is funded by Huawei Noah’s Ark Lab.
7. References


tic and textual data augmentation for improved ASR of code-


[16] Y. Li and P. Fung, “Language modeling with functional head con-