Improved Neural Machine Translation with Chinese Phonologic Features

Jian Yang[1], Shuangzhi Wu[2], Dongdong Zhang[3], Zhoujun Li[1(✉)], and Ming Zhou[3]

[1] Beihang University, Beijing, China
{yangjian123,lizj}@buaa.edu.cn
[2] Harbin Institute of Technology, Harbin, China
v-shuawu@microsoft.com
[3] Microsoft Researcher Asian, Beijing, China
{dozhang,mingzhou}@microsoft.com

Abstract. Chinese phonologic features play an important role not only in the sentence pronunciation but also in the construction of a native Chinese sentence. To improve the machine translation performance, in this paper we propose a novel phonology-aware neural machine translation (PA-NMT) model where Chinese phonologic features are leveraged for translation tasks with Chinese as the target. A separate recurrent neural network (RNN) is constructed in NMT framework to exploit Chinese phonologic features to help facilitate the generation of more native Chinese expressions. We conduct experiments on two translation tasks: English-to-Chinese and Japanese-to-Chinese tasks. Experimental results show that the proposed method significantly outperforms state-of-the-art baselines on these two tasks.

Keywords: Neural Machine Translation · Chinese phonology

1 Introduction

Neural Machine Translation (NMT) with the attention-based encoder-decoder framework [2] has been proved to be the most effective approach to machine translation tasks on many language pairs [2,13,21,24]. In a conventional NMT model, an encoder reads in source sentences of variable lengths, and transforms them into sequences of intermediate hidden vector representations. With weighted attention operations, the hidden vectors are combined and fed to the decoder to generate target translations.

There have been much work to improve the performance of NMT models, such as exploring novel network architectures [7,22] and introducing prior knowledge of syntax information [4,6,12,23]. As the translation accuracy of NMT increases along with new algorithms and models proposed, it still suffers from the challenge of generating idiomatic and native translation expressions on target languages. Intuitively, native expressions may relate to phonologic knowledge of a

M. Zhang et al. (Eds.): NLPCC 2018, LNAI 11108, pp. 303–315, 2018.
https://doi.org/10.1007/978-3-319-99495-6_26

language beyond the surface form of words. In terms of Chinese, linguists pointed that Chinese phonologic features play an important role in both the sentence pronunciation and the construction of native Chinese sentences [25]. For instance, in the translation example of Fig. 1, the meaning of the verb phrase 'raise money' can be literally represented in Chinese as "筹集 (raise) 钱 (money)", "筹集 (raise) 资金 (money)" or "筹 (raise) 钱 (money)". But the last two Chinese expressions are more native than the first one. The reason is that, from Chinese phonologic perspective, the verb-object pair is more common to have the same number of syllables. Therefore, the disyllable-disyllable collocation "筹集 (raise) 资金 (money)" and the monosyllable-monosyllable collocation "筹 (raise) 钱 (money)" acted as verb-object pairs in the references appear more native than the disyllable-monosyllable collocation "筹集 (raise) 钱 (money)" in the NMT baseline. There was previous work applying phonologic features to significantly improve the performance of tasks such as name entity recognition [3]. But Chinese phonologic features have not been explored in translation tasks when Chinese is the target language. In this paper, we propose a novel phonology-aware neural machine translation (PA-NMT) model where Chinese phonologic features are taken into account for translation tasks with Chinese as target. A PA-NMT model encodes source inputs with bi-directional RNNs and associates them with target word prediction via attention mechanism as in most NMT models, but it comes with a new decoder which is able to leverage Chinese phonologic features to help facilitate the generation of target Chinese sentences. Chinese phonology is equivalently represented by Chinese Pinyin, which includes syllable structure (the sequence of Chinese words) and intonation. Intonation mainly consists of high-level tone(first tone), rising tone(second tone), low tone(third tone), falling tone(fourth tone) and neutral tone. Our new decoder in PA-NMT consists of two RNNs. One is to generate the sequence of translation words, and the other is to produce the corresponding Chinese phonologic features, which are further used to help the selection of translation candidates from a phonological perspective.

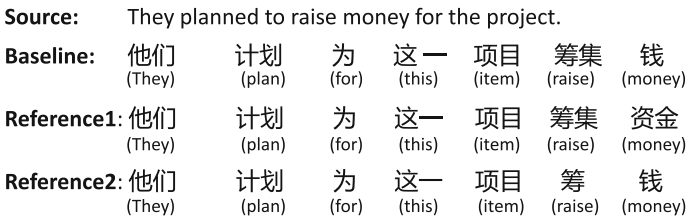

Fig. 1. The meaning of the translation in NMT baseline is correct, but its expression is not native comparing to the references.

We evaluate our method on publicly available data sets with English-Chinese and Japanese-Chinese translation tasks. Experimental results show that our model significantly improves translation accuracy over the conventional NMT baseline systems.

The major contribution of our work is two folds:

(1) We propose a new PA-NMT model to leverage Chinese phonologic features. Our PA-NMT can encode target phonologic features and use them to help rank the translation candidates, so that the translation results of PA-NMT can be more native and accurate.
(2) Our PA-NMT model can achieve high quality results on both English-Chinese and Japanese-Chinese translation tasks, on publicly available data sets.

2 Background: Neural Machine Translation

Neural Machine Translation (NMT) is an end-to-end paradigm [2,7,22] which directly models the conditional translation probability $P(Y|X)$ of the source sentence X and the target translation Y. It usually consists of three parts: the encoder, attention mechanism and the decoder.

For the RNN based NMT model, the RNN encoder bidirectionally encodes the source sentence into a sequence of context vectors $H = h_1, h_2, h_3, ..., h_m$, where m is the source sentence length and $h_i = [\boldsymbol{h}_i, h_i]$, $\boldsymbol{h}_i$ and h_i are calculated by two RNNs from left-to-right and right-to-left respectively. The RNN can be a Gated Recurrent Unit (GRU) [5] or a Long Short-Term Memory (LSTM) [9] in practice. In this paper, we use GRU for all RNNs.

Based on the encoder states, the decoder generates target translations with length n word by word with probability

$$P(Y|X) = \prod_{j=1}^{n} P(y_j|y_{<j}, H) \tag{1}$$

The probability $P(y_j|y_{<j}, H)$ for the jth target word is computed by

$$P(y_j|y_{<j}, H) = g(s_j, y_{j-1}, c_j) \tag{2}$$

where g is a nonlinear, potentially multi-layered, function that outputs the probability of y_j, s_j is the j-th hidden state of decoder RNN, computed by

$$s_j = f_{RNN}(y_{j-1}, s_{j-1}, c_j)$$

c_j is the source context which is calculated by the attention mechanism. The attention mechanism is proposed to softly align each decoder state with the encoder states, where the attention score a_{jk} is computed to explicitly quantify how much each source word contributes to the target word by the following equations,

$$a_{jk} = \frac{\exp(e_{jk})}{\sum_{d=1}^{m} \exp(e_{jd})} \tag{3}$$

The calculation for e_{jk} can be in several ways [14], in this paper we compute e_{jk} by

$$e_{jk} = v_a^T \tanh(W_a s_{j-1} + U_a h_k) \tag{4}$$

where v_a, W_a, U_a are the weight matrix. The final source context c_j is the weighted sum of all encoder states

$$c_j = \sum_{k=1}^{n} a_{jk} h_k. \tag{5}$$

3 Phonology-Aware Neural Machine Translation Model

A phonology-aware neural machine translation (PA-NMT) model is an extension to the conventional NMT model augmented with Chinese phonologic features. The Chinese phonology is equivalently represented by Chinese Pinyin, so we model the phonologic features by Chinese Pinyin associated with tones. For a Chinese sentence, it has a corresponding Pinyin sequence with the same length. $\mathcal{P}_{T-2}, \mathcal{P}_{T-1}$Given a source sentence $X = x_1, x_2, .., x_m$, its target translation $Y = y_1, y_2, .., y_n$ and Y's Pinyin sequence $\mathcal{P} = \mathcal{P}_1, \mathcal{P}_2, .., \mathcal{P}_n$, the goal of our model is to use $\mathcal{P}$ to help the generation of Y. Figure 2 sketches the global overview of our PA-NMT model. Our model first encode source words in the conventional way as described in Sect. 2. In decoder, we use two recurrent neural networks (RNN) to model phonologic features and generate target words. At each timestep, the word RNN generates a list of translation candidates and the phonetic RNN helps to re-score the translation candidates based on phonologic features. Specially, we leverage two attention models for the two decoder RNNs. One is for modeling the phonetic features, the other is used for target word generation.

4 Model Encoder

Our encoder follows the standard RNN encoder [2] (left part in Fig. 2), which bidirectionally reads the input sequence and generate a sequence of context vectors $H = h_1, h_2, h_3, ..., h_m$, where m is the source sentence length and $h_i = [\boldsymbol{h}_i, h_i]$, $\boldsymbol{h}_i$ and h_i are calculated by two RNNs as described in Sect. 2.

4.1 Phonetic-Aware Decoder

Unlike the standard decoder [2], we use phonetic RNN to read Pinyins of history words and rescore current translation candidates as shown in Fig. 2 right top part. The right bottom part is a standard RNN decoder. We map the word dictionary to a Pinyin dictionary one by one, thus the words and Pinyins can

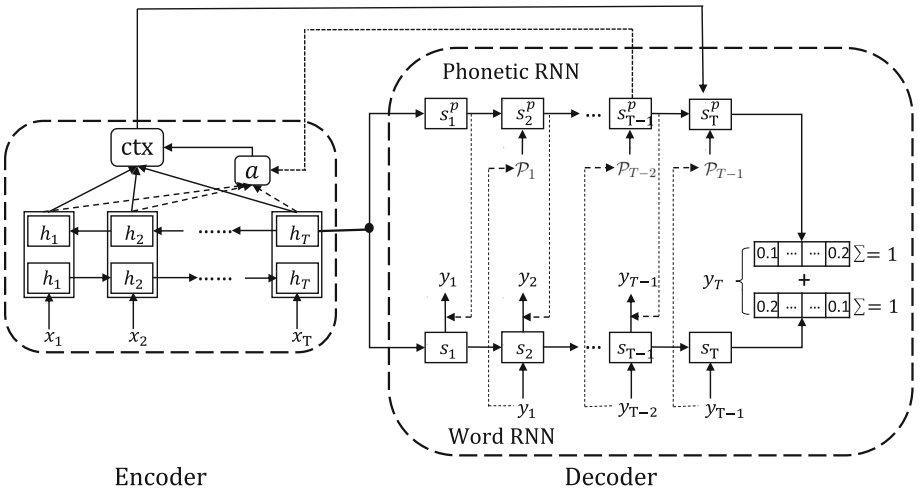

Fig. 2. Overview of PA-NMT model. The phonetic RNN only takes the Pinyin of history words as input and helps re-score the translation candidates. The word RNN is a standard RNN decoder with attention model which is omitted in the figure to simplify readability.

be aligned in their dictionaries. During decoding we force the phonetic RNN to read Pinyins of previous words generated by the word RNN,

$$y_i^{\mathcal{P}} = \sigma_{\mathcal{P}}(y_{i-1}) \tag{6}$$

where $\sigma_{\mathcal{P}}$ is the function which map predicted word to Pinyin which has the most probability. Thus the decoding procedure of the two RNNs can be aligned.

Although both RNNs have separate parameters, word RNN is in a coherent feature with phonetic RNN. By denoting the hidden state of phonetic decoder as $s_i^{\mathcal{P}}$, the calculation in phonetic RNN is as follows,

$$s_i^{\mathcal{P}} = GRU^{\mathcal{P}}(s_{i-1}^{\mathcal{P}}, y_{i-1}^{\mathcal{P}}, c_i^{\mathcal{P}}) \tag{7}$$

where $c_i^{\mathcal{P}}$ is the source context vector and $y_{i-1}^{\mathcal{P}}$ is the Pinyin of previous word y_{i-1}. The context vector $c_i^{\mathcal{P}}$ depends on the source states $[h_1, ..., h_m]$ and is calculated by the attention model,

$$c_i^{\mathcal{P}} = \sum_{j=1}^{last} a_{ij}^{\mathcal{P}} \cdot h_j \tag{8}$$

The weight $a_{ij}^{\mathcal{P}}$ of each annotation h_j is computed by

$$a_{ij}^{\mathcal{P}} = \frac{\exp(e_{ij})}{\sum_{k=1}^{last} \exp(e_{ik})} \tag{9}$$

where e_{ij} is computed as

$$e_{ij}^{\mathcal{P}} = v_{\mathcal{P}}^{T} \tanh(W_{\mathcal{P}}[s_{i-1}^{\mathcal{P}}; h_j]) \tag{10}$$

With $s_i^{\mathcal{P}}$ and $c_i^{\mathcal{P}}$, the phonetic can calculate a probability list by softmax as

$$p(\mathcal{P}_i|\mathcal{P}_1, ..., \mathcal{P}_{i-1}, x) = g(\mathcal{P}_{i-1}, s_i, c_i^{\mathcal{P}}) \tag{11}$$

We rescore the prediction of a new word by adding the log probability of the two softmax list. Thus the score of the new translation candidate y_i is

$$\text{score} = p(y_j|y_{<j}, X) + \alpha \log(\mathcal{P}_j|\mathcal{P}_{<j}, X) \tag{12}$$

where $\mathcal{P}_j$ is the corresponding Pinyin for the Chinese word y_j, and α is a hyper-parameter which control the importance of Chinese phonologic decoder. We set α to 0.5 in experiments. We find that when α is too big, it will make model worse. When α approximates zero, our model fails to extract phonological feature. In order to better preserve the Pinyin information, we use two attention parameters to store the Chinese character and pinyin alignment information respectively.

In our model, one RNN extracting phonological features interacts with another RNN in two aspects. For every timestep, word RNN generates next input of phonetic RNN which is converted from word to Pinyin to keep consistent with phonetic RNN. While predicting next word, phonetic RNN evaluate results of word RNN to generate final predicted candidates.

4.2 Chinese Polyphone Disambiguation

When mapping Chinese words to Pinyins, the major problem is that there could be multiple polyphone candidates for one Chinese word. Given a Chinese sentence, its Pinyin sequence expression is usually deterministic based on the context of the whole sentence. In our work, to align the Pinyin sequence and word sequence, the Pinyins are generated from the words with context-free information. To make disambiguation of Pinyin generation, we heuristically map each Chinese word to the Pinyin with the highest probability in terms of statistics over a Chinese monolingual corpus of 20 million sentences.

4.3 Model Training

Different form the objective function to train the conventional NMT model, for our joint PA-NMT model, we use the sum of log-likelihoods of word sequence and Pinyin sequence as our objective function:

$$J(\theta) = \sum_{(X,Y,\mathcal{P})\in D} \log\ P(\mathcal{P}|X) + \log\ P(Y|X) \tag{13}$$

Thus, our training data format is (source sentence, (target Chinese sentence, target Chinese Pinyin)). In this way, we incorporate the phonological features into Chinese sentence generation.

5 Experiment

5.1 Setup

In the English-Chinese task, we use a subset from LDC corpus[1] which has around 2.6M sentence pairs from News domain. We use NIST 2008 as testset which has 4 references for each source sentence. We also make several other testsets by reversing the direction of Chinese-English sets NIST 2003, NIST 2005 and NIST 2012, as these sets all have four English reference, we just use the first reference as English source sentence. We use the WMT2009 English-Chinese set for development.

In the Japanese-Chinese task, we use 2.87M sentence pairs from ASPEC Japanese-Chinese corpus [15][2]. The development data contains 1, 784 sentences, and the test data contains 1, 812 sentences with single reference per source sentence. Both source and target language are tokenized with our in-house tools.

For the training data of target Chinese sentences in both the English-Chinese task and the Japanese-Chinese task, we covert them into Chinese Pinyin using our in-house implemented tool based on a statistical translation model with the accuracy of above 90%.

In the neural network training, the vocabulary size is limited to 30 K high frequent words for both source and target languages. All low frequent words are normalized into a special token `unk` and post-processed by following the work in [14]. The size of word embedding and transition action embedding is set to 512. The dimensions of the hidden states for all RNNs are set to 1024. All model parameters are initialized randomly with Gaussian distribution [8] and trained on a NVIDIA Tesla 1080 GPU. The stochastic gradient descent (SGD) algorithm is used to tune parameters with a learning rate of 1.0. The batch size is set to 128. In the update procedure, Adadelta [26] algorithm is used to automatically adapt the learning rate. The beam sizes for both word prediction and transition action prediction are set to 8 in decoding.

The baselines in our experiments is a neural translation system, denoted by RNNsearch which is an in-house implementation of the attention-based neural machine translation model [2] using the same parameter settings as our PA-NMT model. The evaluation results are reported with the word level and character level case-insensitive IBM BLEU-4 [17] denoted as **word-BLEU** and **char-BLEU** respectively. A statistical significance test is performed using the bootstrap resampling method proposed by [11] with a 95% confidence level.

5.2 Evaluation Results

We first evaluate our method on the English-Chinese translation task. The evaluation results over all test sets against baselines are listed in bottom part

[1] LDC2002E17, LDC2002E18, LDC2003E07, LDC2003E14, LDC2005E83, LDC2005T06, LDC2005T10, LDC2006E17, LDC2006E26, LDC2006E34, LDC2006E85, LDC2006E92, LDC2006T06, LDC2004T08, LDC2005T10.

[2] http://orchid.kuee.kyoto-u.ac.jp/ASPEC/.

Table 1. Evaluation results on English-Chinese and Japanese-Chinese translation tasks with word-BLEU% and char-BLEU% metrics. The "Average" row in the English-Chinese part refers to the averaged result of all test sets. The numbers in bold indicate statistically significant difference ($p < 0.05$) from baselines.

Japanese-Chinese				
	RNNsearch		PA-NMT	
	word-BLEU	char-BLEU	word-BLEU	char-BLEU
dev	33.41	44.49	**34.03**	**45.23**
devtest	33.38	44.35	**34.26**	**45.26**
test	33.53	44.57	**34.19**	**45.26**
English-Chinese				
	RNNsearch		PA-NMT	
	word-BLEU	char-BLEU	word-BLEU	char-BLEU
NIST2003	18.38	30.79	**19.34**	**32.07**
NIST2005	17.02	28.87	**17.90**	**30.17**
NIST2008	22.71	34.30	**23.94**	**35.10**
NIST2012	13.51	23.33	**14.37**	**24.52**
Avg.	17.91	29.32	18.89	30.47

of Table 1. From the table, our PA-NMT outperforms RNNsearch on all the test sets on both word- and char- BLEU scores, where our model surpasses the baseline most on NIST 2008 set with 1.23 and 0.80 more scores on the two metrics. In terms of the average word-BLEU scores, our model outperforms the baseline by 0.98 BLEU points. And on the average char-BLEU scores, our model also outperforms the baseline by 1.14 BLEU points which shows our proposed phonology-aware NMT model performs much better than traditional sequence-to-sequence NMT model.

We also report results on the Japanese-Chinese translation task. The top part of Table 1 shows the comparison results with the evaluation metrics of word- and char- BLEU. From the table, our method outperforms the NMT baseline on the three datasets in terms of both word- and char- BLEUs.

5.3 Case Study

We sampled some translation examples from the test sets to make case study of how our method can improve the English-Chinese translation task. In the examples in Table 2, there are merely two phonology aspects we investigated: auxiliary word and syllable repetition.

Auxiliary word Structural auxiliary words have almost no actual grammatical meaning but play a role in the language structure. They just express the sound of the utterance or make the syllable of the language symmetry. In modern Chinese,

Table 2. Translation examples of RNNsearch and our PA-NMT on English-Chinese translation task. RNNsearch fails to generate pure Chinese sentences. Whereas with the help of the phonologic knowledge, our PA-NMT can get much better translations.

source	we cannot leave a heavy burden for later generations : this includes the issue of the rapid expansion of the population
RNNsearch	我们 不能 留给 后人 的 沉重 负担 : 这 包括 人口 的 快速 发展 的 问题 wǒmen búnéng liúgěi hòurén de chénzhòng fùdān : zhè bāokuò rénkǒu de kuàisù fāzhǎn de wèntí
PA-NMT	我们 不能 留给 后人 沉重 负担 : 这 包括 人口 急速 膨胀 的 问题 wǒmen búnéng liúgěi hòurén chénzhòng fùdān : bāokuò rénkǒu jísù péngzhàng de wèntí
reference	我们 不可 为 后代 人 留下 沉重 的 负担 , 包括 人口 急速 膨胀 的 问题 wǒmen búkě wéi hòudài rén liúxià chénzhòng de fùdān , bāokuò rénkǒu jísù péngzhàng de wèntí
source	London underground train derailed , injuring 37 people
RNNsearch	伦敦 地下 火车 出轨 , 伤 人 37 人 lúndūn dìxià huǒchē chūguǐ , shāng rén sānshí qī rén
PA-NMT	伦敦 地下 火车 出轨 事故 造成 37 人 受伤 lúndūn dìxià huǒchē chūguǐ shìgù zàochéng sānshíqī rén shòushāng
reference	伦敦 地铁 列车 出轨 三十七 人 受伤 lúndūn dìtiě lièchē chūguǐ sānshíqī rén shòushāng
source	zhoushan is one of chinese cities suffering from a serious water shortage
RNNsearch	舟山市 是 一个 严重 缺 水 的 城市 之一 zhōushānshì shì yīgè yánzhòng quē shuǐ de chéngshì zhīyī
PA-NMT	舟山 是 中国 城市 缺水 较为 严重的 城市 之一 zhōushān shì zhōngguó chéngshì quēshuǐ jiàowéi yánzhòng de chéngshì zhīyī
reference	舟山市 是 中国 缺 水 较为 严重 的 城市 之一 zhōushān shì zhōngguó quē shuǐ jiàowéi yánzhòng de chéngshì zhīyī

“的 (de)” is one of the important structural auxiliary words. But if the word “的 (de)” in a incorrect position, it will disturb syllables of the whole sentence.

As shown in the first example of Table 2, “我们不能留给后人的沉重负担” can be represented by Pinyin as “wǒmen búnéng liúgěi hòurén de chénzhòng fùdān”. “的 (de)” splits the Pinyin sequence into two parts “wǒmen búnéng liúgěi hòurén de” and “chénzhòng fùdān”. The former part modifies the latter part. It will lead a unbalanced pronunciation for native speakers, because both “wǒmen búnéng liúgěi hòurén de” and “chénzhòng fùdān” have too many syllables which will make the whole sentence unstable and lead difficulty of speaking. However, in the sentence “我们不能留给后人沉重负担 (búnéng liúgěi hòurén chénzhòng fùdān)”, “沉重 (chénzhòng)” modify “负担 (fùdān)”. From semantic perspective, both “wǒmen búnéng liúgěi hòurén de chénzhòng fùdān” and “wǒmen búnéng liúgěi hòurén chénzhòng de fùdān” have the same meaning in

Chinese. “沉重 (chénzhòng)” and “负担 (fùdān)” are both two-syllable word, “沉重 (chénzhòng) 负担 (fùdān)” will be more easy to read. Our model’s translation is “我们 (we) 不能 (cannot) 给 (give) 后人 (later generations) 留下 (leave) 沉重 (heavy) 的负担 (burden)” and it’s Pinyin is “wǒmen búnéng gěi hòuràn liúxià chénzhòng fùdān”. The sentence don’t have “(的)de” in a incorrect position of the whole sentence which ensure keeping the original meaning unchanged and make pronunciation of sentence more fluent and authentic in Chinese.

Syllable repetition Syllable repetition also affect the rhythm of sentences. In second example, “injuring 37 people” is translated into “ 伤人 (injuring) 人 (people)” in RNNsearch and “造成 (make) 人 (people) 受伤 (injured)” in our model. We are not used to speaking continuous “ren” in Chinese because it can break the rhythm of whole sentence. Pinyin of “伤人 (injuring) 人 (people)” is “shāngrén sānshíqī rén”. The Pinyin sequence is spoken as two parts “shāngrén” which is word and “sānshíqī rén” which has four syllables. Two adjacent sequence both have “人 (ren)” as their ending. On the one hand, Except some special usage, we are not used to it. Generally speaking, we can speak “造成 (make) 人 (people) 受伤 (injured)” or “人 (people) 受伤 (injured)” which avoid continuously use the same monosyllable word in our daily. On the other hand, “造成 (make) 人 (people) 受伤 (injured)” has three parts “造成 (make)” , “人 (people)” and “受伤 (injured)”. They all have similar number of syllables. We are accustomed to using this style which make the whole sentence are more rhythmic.

In the last example, we can see the RNNsearch’s translation express the same meaning “一 (one)” twice. But this kind of expression of Chinese is illegal in grammar. The first “一个 (one)” and “之一 (one of)” behind it together is regarded as an adjective and play a role in attribute. The first “一个 (one)” will result in a semantic and phonological repetition in Chinese. The RNNsearch consider more on semantic information. PA-NMT consider both semantic feature and phonological feature. In most real cases, Chinese native speaker rarely express the similar words twice despite that they play different roles in the sentence. Hence, the Chinese sentence only keeps “之一 (one of)” translated by PA-NMT. We can see that PA-NMT avoid the repetition of pronunciation.

These examples mainly reflects in the usage of syllable structure in Chinese. Our model generates more native Chinese sentence with considering the combination of syllables and repetition of several syllables. Because Chinese characters are monosyllables, the syllables in sentences are very important in English-Chinese translation which considers balancing syllables in sentence. Each word only has each syllable. Hence, Pinyin is suitable for model to collect phonological features. By introducing Chinese phonologic feature, our model can learn the both semantic and phonetic information.

6 Related Work

Recently, neural machine translation (NMT) has achieved better performance than SMT in many language pairs [13,16,19,24,27]. A lot of work has been

done to incorporate linguistic knowledge into NMT models [4,6,12,23]. A tree-to-sequence attentional NMT model is proposed in [6] where source-side HPSG tree was used. [4] leveraged the phrase-structure trees as in the Penn Chinese Treebank as prior knowledge for NMT inputs. They proposed a tree-coverage model to let the attention depend on the source-side syntax. In these models, the source dependency structure is used. For the target side, [1] proposed to replace the target sentence with the linearized, lexicalized constituency tree. [23] proposed to jointly learn target translation and dependency parsing.

Many languages like Russian use morphological information to improve translation quality in recent work [20]. Chinese linguistic features have been leveraged to help NLP tasks. For example, Chinese radicals were used as additional features to improve machine translation [10,18]. Chinese phonologic features were explored to address named entity recognition problem [3]. Different from those work, in this paper we propose to involve Chinese target phonetic features into NMT model to help translate pure and native Chinese sentences.

7 Conclusion

In this paper, we propose a novel phonology-aware neural machine translation (PA-NMT) model where Chinese phonologic features are used for translation tasks with Chinese as target. Our model encodes these features in the NMT decoder by another recurrent neural network (RNN), aiming to help facilitate the generation of target Chinese sentences. Our method tries to collect and use phonological features to optimize language model which is different from RNNsearch. Experimental results show that our method can boost the translation generation and achieve significant improvements on the translation quality of NMT systems. Along this research direction, in future work we will try to integrate other prior knowledge, such as semantic information, into NMT systems.

Acknowledgments. This work was supported in part by the Natural Science Foundation of China (Grand Nos. U1636211,61672081,61370126), and Beijing Advanced Innovation Center for Imaging Technology (No. BAICIT-2016001) and National Key R&D Program of China (No. 2016QY04W0802).

References

1. Aharoni, R., Goldberg, Y.: Towards string-to-tree neural machine translation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 132–140. Association for Computational Linguistics, Vancouver, Canada, July 2017. http://aclweb.org/anthology/P17-2021
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: ICLR 2015 (2015)
3. Bharadwaj, A., Mortensen, D., Dyer, C., Carbonell, J.: Phonologically aware neural model for named entity recognition in low resource transfer settings. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1462–1472 (2016)

4. Chen, H., Huang, S., Chiang, D., Chen, J.: Improved neural machine translation with a syntax-aware encoder and decoder. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1936–1945. Association for Computational Linguistics, Vancouver, Canada, July 2017. http://aclweb.org/anthology/P17-1177
5. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of ENMLP 2014, October 2014
6. Eriguchi, A., Hashimoto, K., Tsuruoka, Y.: Tree-to-sequence attentional neural machine translation. In: Proceedings of ACL 2016, August 2016
7. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning. arXiv preprint arXiv:1705.03122 (2017)
8. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. Aistats **9**, 249–256 (2010)
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
10. Zhang, J., Matsumoto, T.: Improving character-level Japanese-Chinese neural machine translation with radicals as an additional input feature (2003)
11. Koehn, P.: Statistical significance tests for machine translation evaluation. In: EMNLP, pp. 388–395. Citeseer (2004)
12. Li, J., Xiong, D., Tu, Z., Zhu, M., Zhang, M., Zhou, G.: Modeling source syntax for neural machine translation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 688–697. Association for Computational Linguistics, Vancouver, Canada, July 2017. http://aclweb.org/anthology/P17-1064
13. Luong, T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Proceedings of EMNLP 2015, September 2015
14. Luong, T., Sutskever, I., Le, Q., Vinyals, O., Zaremba, W.: Addressing the rare word problem in neural machine translation. In: Proceedings of ACL 2015, July 2015
15. Nakazawa, T., et al.: ASPEC: Asian scientific paper excerpt corpus. In: Chair, N.C.C., et al. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016), pp. 2204–2208. European Language Resources Association (ELRA), Portoroz, Slovenia, May 2016
16. Neubig, G.: Lexicons and minimum risk training for neural machine translation: NAIST-CMU at WAT2016. In: Proceedings of the 3rd Workshop on Asian Translation (WAT2016), Osaka, Japan, December 2016
17. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of ACL 2002 (2002)
18. Kuang, S., Han, L.: Apply Chinese radicals into neural machine translation: deeper than character level
19. Shen, S., et al.: Minimum risk training for neural machine translation. In: Proceedings of ACL 2016, August 2016
20. Song, K., Zhang, Y., Zhang, M., Luo, W.: Improved English to Russian translation by neural suffix prediction. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (2018)
21. Tu, Z., Lu, Z., Liu, Y., Liu, X., Li, H.: Modeling coverage for neural machine translation. In: Proceedings of ACL 2016, August 2016
22. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 6000–6010 (2017)

23. Wu, S., Zhang, D., Yang, N., Li, M., Zhou, M.: Sequence-to-dependency neural machine translation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 698–707. Association for Computational Linguistics, Vancouver, Canada, July 2017. http://aclweb.org/anthology/P17-1065
24. Wu, Y., et al.: Google's neural machine translation system: bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)
25. Xia, L.: A brief discussion on phonology and rhythm beauty in English-Chinese translation (2003). https://wenku.baidu.com/view/c3666404a200a6c30c22590102020740be1ecd76.html
26. Zeiler, M.D.: ADADELTA: an adaptive learning rate method. arXiv preprint arXiv:1212.5701 (2012)
27. Zhang, B., Xiong, D., Su, J., Duan, H., Zhang, M.: Variational neural machine translation. In: Proceedings of EMNLP 2016, November 2016