Open Vocabulary Learning for Neural Chinese Pinyin IME

Zhuosheng Zhang, Yafang Huang, Hai Zhao
Shanghai Jiao Tong University
{zhangzs, huangyafang}@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn

Abstract

Pinyin-to-character (P2C) conversion is the core component of pinyin-based Chinese input method engine (IME). However, the conversion is seriously compromised by the ambiguities of Chinese characters corresponding to pinyin as well as the predefined fixed vocabularies. To alleviate such inconveniences, we propose a neural P2C conversion model augmented by an online updated vocabulary with a sampling mechanism to support open vocabulary learning during IME working. Our experiments show that the proposed method outperforms commercial IMEs and state-of-the-art traditional models on standard corpus and true inputting history dataset in terms of multiple metrics and thus the online updated vocabulary indeed helps our IME effectively follows user inputting behavior.

Introduction

Motivation: Converting pinyin to Chinese characters is the most basic module of all pinyin-based IMEs. It is natural to regard the Pinyin-to-Character (P2C) conversion as a machine translation between two different languages, pinyin sequences and Chinese character sequences.

Challenge: too much ambiguity mapping pinyin syllable to character. Pinyin IME may benefit from decoding longer pinyin sequence for more efficient inputting. When a given pinyin sequence becomes longer, the list of the corresponding legal character sequences will significantly reduce.

Pinyin seq. consists of **1** syllable	bei	jing	huan	ying	ni
	被	敬	环	英	**你**
	北	静	换	颖	睨
	呗	井	还	**迎**	逆
	杯	**京**	幻	影	拟
	背	经	**欢**	应	尼
Pinyin seq. consists of **2** syllables	bei_jing		huan_ying		ni
	北京		幻影		你
	背景		欢迎		你
Pinyin seq. consists of **5** syllables	bei_jing_huan_ying_ni				
	北京欢迎你				

Table 1: The shorter the pinyin sequence is, the more character sequences will be mapped.

Observation: User's inputting style may change from time to time, let alone diverse user may input quite diverse contents, which makes a predefined fixed vocabulary can never be sufficient.

Solution: an open vocabulary learning framework,

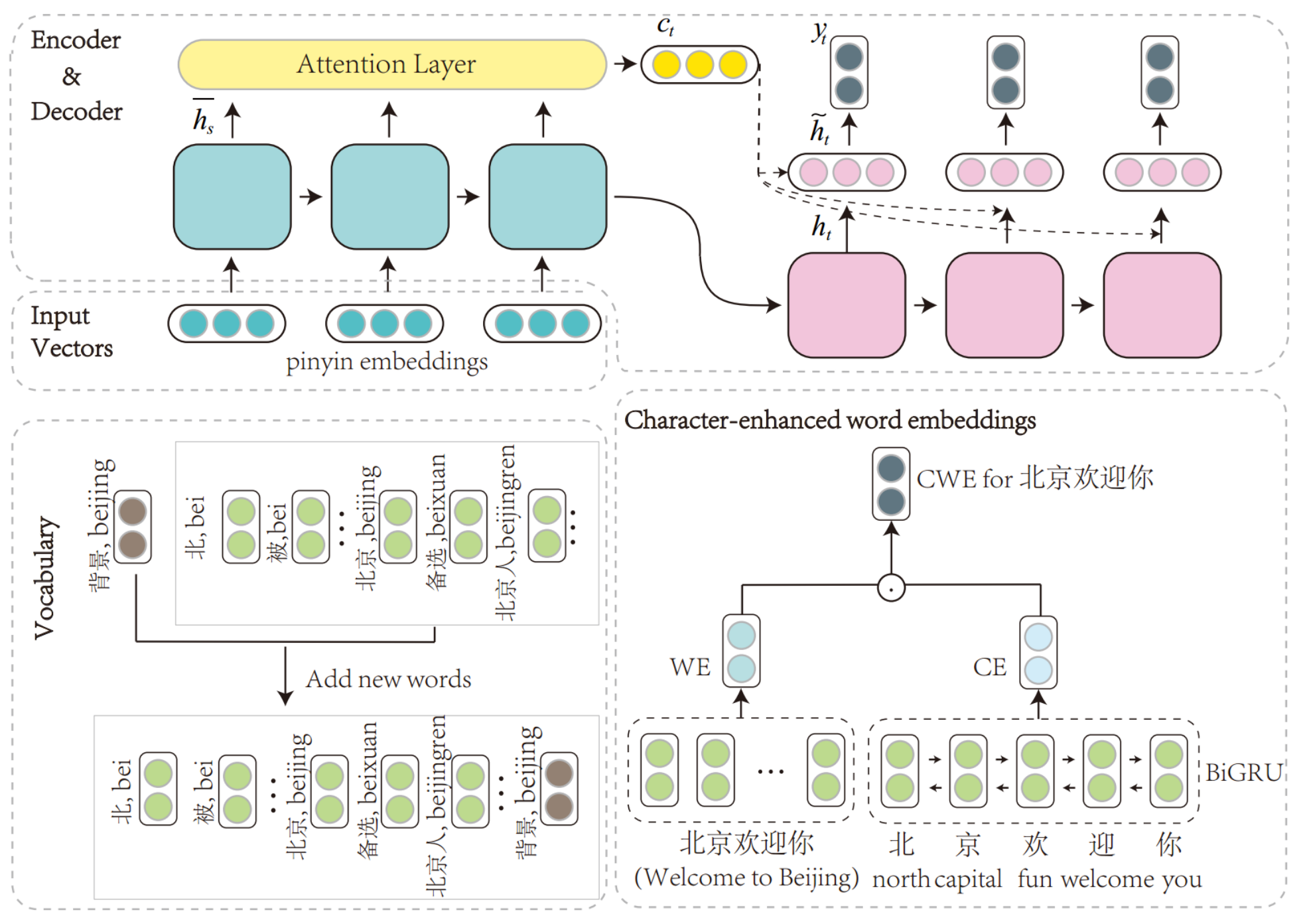
- An online sequence-to-sequence model for P2C
- A sampling mechanism utilizing our online updated vocabulary to enhance the conversion accuracy of IMEs as well as speed up the decoding procedure.

In detail,

- A character-enhanced word embedding (CWE) mechanism is proposed to obtain fine-grained word representation and pick a very small target vocabulary for each sentence.
- Every time the user makes a selection contradicted the prediction given by the P2C conversion module, the module will update the vocabulary accordingly.

Method

The core of P2C is the encoder-decoder framework. The encoder is a BiLSTM network with global attention mechanism.



Character-enhanced Word Embedding

We adopt a hybrid mechanism to balance both words and characters representation, namely, Character-enhanced Word Embedding (CWE).

In the beginning, we keep an initial vocabulary with the most frequent words. The words inside the vocabulary are represented as enhanced-embedding, and those outside the list are computed from character embeddings.

Online P2C Learning with Vocabulary Adaptation

Aim: track the continuous change of users' inputting contents.

Method: The updating procedure introduces new words by comparing the user's choice and IME's top-1 prediction. The longest mismatch n-gram characters will be added as new word.

Algorithm 1 Online Vocabulary Updating Algorithm

Input:

- Vocabulary: $V = \{(Py_i, Ch_i)|i = 1, 2, 3, \cdots\}$;
- Input pinyin sequence: $Py = \{py_i|i = 1, 2, 3, \cdots\}$;
- IME predicted top-1 character sequence: $Cm = \{cm_i|i = 1, 2, 3, \cdots\}$;
- User choosing character sequence: $Cu = \{cu_i|i = 1, 2, 3, \cdots\}$.

Output:

- The Updated Vocabulary: $\hat{V}$.

```
▷ Adding new words
for $n = 6$ to 2 do
    Compare $n$-gram of $Cu$ and $Cm$
    if Mismatch $Ch$ is found // both the first and last characters are different at least then
        if $Ch$ is not in $\hat{V}$ then
            $V = V \cup \{Ch\}$
        end if
    end if
    if no mismatch is found then
        break
    end if
end for
return $\hat{V}$;
```

Target Vocabulary Selection

We maintain a separate and small vocabulary for each sentence so that we only need to compute the probability distribution over a small vocabulary for each sentence.

Datasets and Metrics

- Dataset: the People's Daily corpus and the TouchPal corpus. The corpora and our codes are available at https://github.com/cooelf/OpenIME

		Chinese	Pinyin
PD	# MIUs	5.04M	
	# Word	24.7M	24.7M
	# Vocab	54.3K	41.1K
	# Target Vocab (train)	2309	-
	# Target Vocab (dec)	2168	-
TP	# MIUs	689.6K	
	# Word	4.1M	4.1M
	# Vocab	27.7K	20.2K
	# Target Vocab (train)	2020	-
	# Target Vocab (dec)	2009	-

- Evaluation metrics: Maximum Input Unit (MIU) Accuracy and KeyStroke Score (KySS).

Result

Baseline systems:

- Google IME 2
- Offline and Online models for Word Acquisition (OMWA, On-OMWA)

System	ED	PD			TP		
		Top1	Top5	Top10	Top1	Top5	Top10
Existing P2C							
Google IME		70.9	78.3	**82.3**	57.5	63.8	69.3
OMWA		55.0	63.7	70.2	19.7	24.8	27.7
On-OMWA		64.4	72.9	77.9	57.1	71.1	80.9
Our P2C							
Base P2C	200	53.2	64.7	70.3	46.8	68.8	75.7
On-P2C	200	68.1	77.3	78.2	69.8	88.7	89.3
On-P2C (bi)	200	70.5	79.8	80.1	71.0	89.2	89.5
On-P2C (bi)	300	70.8	**80.5**	81.2	**71.9**	89.6	**90.6**
On-P2C (bi)	400	**71.3**	80.1	**81.3**	71.7	**89.7**	90.3
On-P2C (bi)	500	69.9	78.2	81.0	70.7	89.2	89.8

Results:

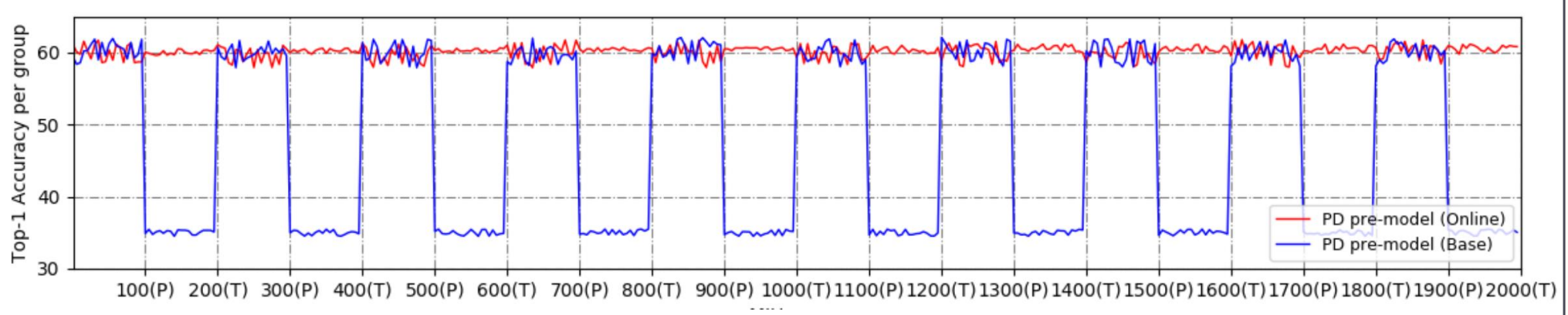
- On the People's Daily corpus, our online model (On-P2C) outperforms the best model in (Zhang et al., 2017) by +3.72% top-1 MIU accuracy.
- The +14.94 improvement over the base P2C conversion module demonstrates that online learning vocabulary is effective.

Analysis

Effects of Online Updated Vocabulary

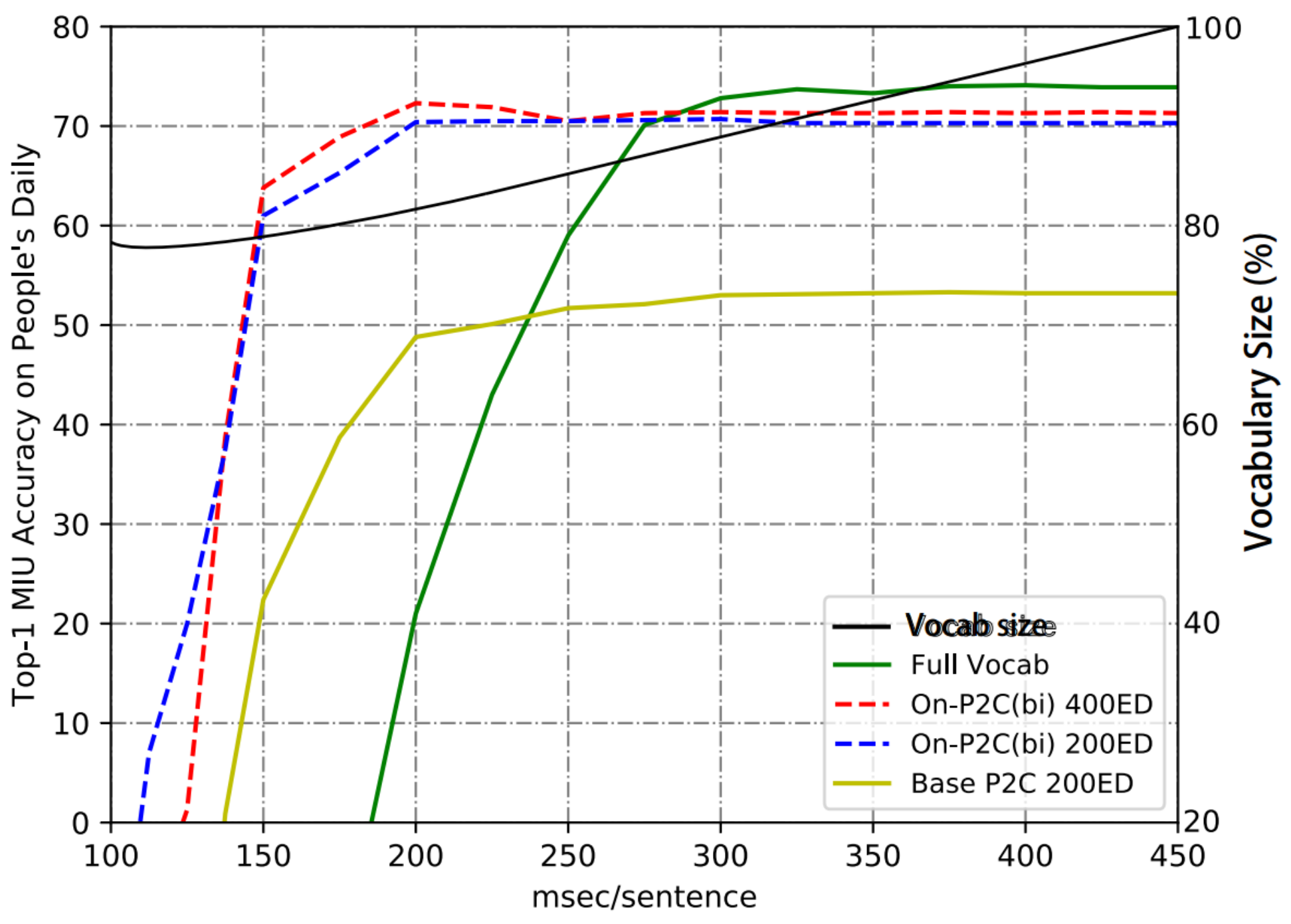
- Models with online vocabulary updating significantly outperform those without updating.
- Online P2C distinctly adapts the corpus change at the joint part. On the contrary, the base P2C which works offline performs stably only on its in-domain segments.

Models	the People's Daily	TouchPal
Google IME	0.7535	0.6465
OMWA	0.6496	0.4489
On-OMWA	0.7115	0.7226
Base P2C	0.6922	0.7910
On-P2C	**0.8301**	**0.8962**



Effects of Vocabulary Selection

- The accuracies nearly do not get decreased with high enough decoding speed when only taking 88.9% full vocabulary in our system.



Effects of Word Filtering for CWE building

- Pure word-level representation is more efficient for P2C tasks than character-level
- Omitting partial low-frequency word is beneficial in establishing word-level embedding.

Filter Ratio	0	0.3	0.6	0.9	1.0
Top-5 Accuracy(valid set)	66.4	68.3	84.3	89.7	87.5
Top-5 Accuracy(test set)	66.3	68.1	83.9	**89.6**	87.1