Moon IME: Neural-based Chinese Pinyin Aided Input Method with Customizable Association

Yafang Huang[1,2], **Zuchao Li**[1,2], **Zhuosheng Zhang**[1,2], **Hai Zhao**[1,2,*],
[1]Department of Computer Science and Engineering,
Shanghai Jiao Tong University, Shanghai, 200240, China
[2]Key Laboratory of Shanghai Education Commission for Intelligent Interaction
and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China
huangyafang@sjtu.edu.cn, charlee@sjtu.edu.cn zhangzs@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn *

Abstract

Chinese pinyin input method engine (IME) lets user conveniently input Chinese into a computer by typing pinyin through the common keyboard. In addition to offering high conversion quality, modern pinyin IME is supposed to aid user input with extended association function. However, existing solutions for such functions are roughly based on oversimplified matching algorithms at word-level, whose resulting products provide limited extension associated with user inputs. This work presents the Moon IME, a pinyin IME that integrates the attention-based neural machine translation (NMT) model and Information Retrieval (IR) to offer amusive and customizable association ability. The released IME is implemented on Windows via text services framework.

1 Introduction

Pinyin is the official romanization representation for Chinese and pinyin-to-character (P2C) which concerts the inputted pinyin sequence to Chinese character sequence is the core module of all pinyin based IMEs. Previous works in kinds of literature only focus on pinyin to the character itself, paying less attention to user experience with associative advances, let along predictive typing or automatic completion. However, more agile association outputs from IME predication may undoubtedly lead to incomparable user typing experience, which motivates this work.

Modern IMEs are supposed to extend P2C with association functions that additionally predict the next series of characters that the user is attempting to enter. Such IME extended capacity can be generally fallen into two categories: auto-completion and follow-up prediction. The former will look up all possible phrases that might match the user input even though the input is incomplete. For example, when receiving a pinyin syllable "bei", auto-completion module will predict "北京" (beijing, Beijing) or "背景" (beijing, Background) as a word-level candidate. The second scenario is when a user completes entering a set of words, in which case the IME will present appropriate collocations for the user to choose. For example, after the user selects "北京" (Beijing) from the candidate list in the above example, the IME will show a list of collocations that follows the word Beijing, such as "市" (city), "奥运会" (Olympics).

This paper presents the Moon IME, a pinyin IME engine with an association cloud platform, which integrates the attention-based neural machine translation (NMT) model with diverse associations to enable customizable and amusive user typing experience.

Compared to its existing counterparts, Moon IME has extraordinarily offered the following promising advantages:

• It is the first attempt that adopts attentive NMT method to achieve P2C conversion in both IME research and engineering.

• It provides a general association cloud platform which contains follow-up-prediction and machine translation module for typing assistance.

• With an information retrieval based module, it realizes fast and effective auto-completion which can help users type sentences in a more convenient and efficient manner.

* Corresponding author. This paper was partially supported by National Key Research and Development Program of China (No. 2017YFB0304100), National Natural Science Foundation of China (No. 61672343 and No. 61733011), Key Project of National Society Science Foundation of China (No. 15-ZDA041), The Art and Science Interdisciplinary Funds of Shanghai Jiao Tong University (No. 14JCRZ04).

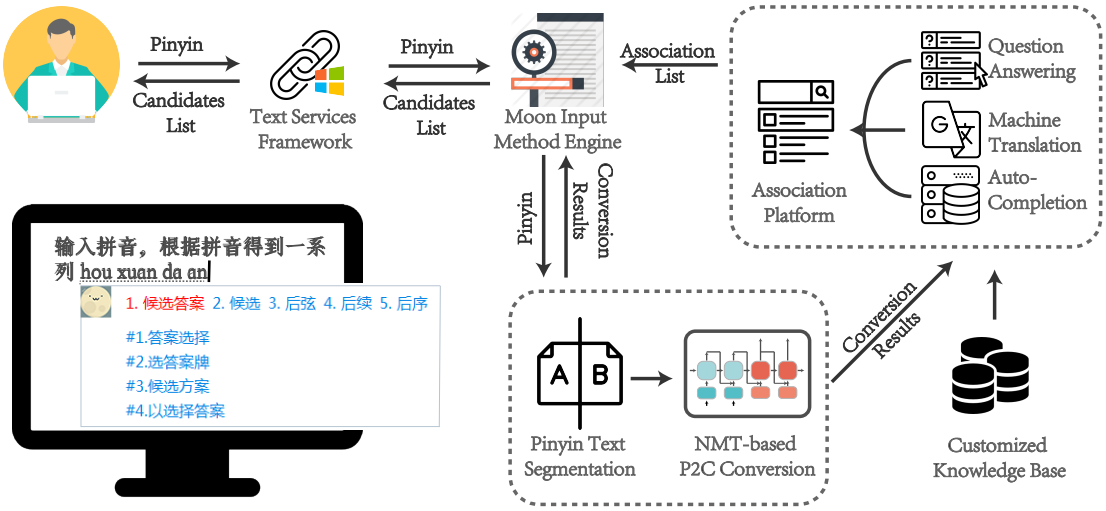

Figure 1: Architecture of the proposed Moon IME.

• With a powerful customizable design, the association cloud platform can be adapted to any specific domains such as the fields of law and medicine which contain complex specialized terms.

The rest of the paper is organized as follows: Section 2 demonstrates the details of our system. Section 3 presents the feature functions of our realized IME. Some related works are introduced in Section 4. Section 5 concludes this paper.

2 System Details

Figure 1 illustrates the architecture of Moon IME. The Moon IME is based on Windows Text Services Framework (TSF) [1]. Our Moon IME extends the Open-source projects PIME[2] with three main components: a) pinyin text segmentation, b) P2C conversion module, c) IR-based association module. The nub of our work is realizing an engine to stably convert pinyin to Chinese as well as giving reasonable association lists.

2.1 Input Method Engine

Pinyin Segmentation For a convenient reference, hereafter a *character* in pinyin also refers to an independent syllable in the case without causing confusion, and *word* means a pinyin syllable sequence with respect to a true Chinese word.

As (Zhang et al., 2017) proves that P2C conversion of 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. Thus, we train our P2C model with segmented corpora. We used baseSeg (Zhao et al., 2006) to segment all text, and finish the training in both word-level and character-level.

NMT-based P2C module Our P2C module is implemented through OpenNMT Toolkit[3] as we formulize P2C as a translation between pinyin and character sequences. Given a pinyin sequence X and a Chinese character sequence Y, the encoder of the P2C model encodes pinyin representation in word-level, and the decoder is to generate the target Chinese sequence which maximizes $P(Y|X)$ using maximum likelihood training.

The encoder is a bi-directional long short-term memory (LSTM) network (Hochreiter and Schmidhuber, 1997). The vectorized inputs are fed to forward LSTM and backward LSTM to obtain the internal features of two directions. The output for each input is the concatenation of the two vectors from both directions: $\overleftrightarrow{\mathbf{h}_t} = \overrightarrow{\mathbf{h}_t} \parallel \overleftarrow{\mathbf{h}_t}$.

Our decoder is based on the global attentional model proposed by (Luong et al., 2015) which takes the hidden states of the encoder into consideration when deriving the context vector. The probability is conditioned on a distinct context vector for each target word. The context vec-

[1] TSF is a system service available as a redistributable for Windows 2000 and later versions of Windows operation system. A TSF text service provides multilingual support and delivers text services such as keyboard processors, handwriting recognition, and speech recognition.

[2] https://github.com/EasyIME/PIME

[3] http://opennmt.net

tor is computed as a weighted sum of previously hidden states. The probability of each candidate word as being the recommended one is predicted using a softmax layer over the inner-product between source and candidate target characters.

Our model is initially trained on two datasets, namely the People's Daily (PD) corpus and Douban (DC) corpus. The former is extracted from the People's Daily from 1992 to 1998 that has word segmentation annotations by Peking University. The DC corpus is created by (Wu et al., 2017) from Chinese open domain conversations. One sentence of the DC corpus contains one complete utterance in a continuous dialogue situation. The statistics of two datasets is shown in Table 1. With character text available, the needed parallel corpus between pinyin and character texts is automatically created following the approach proposed by (Yang et al., 2012).

		Chinese	Pinyin
PD	# MIUs	5.04M	
	# Vocab	54.3K	41.1K
DC	# MIUs	1.00M	
	# Vocab	50.0K	20.3K

Table 1: MIUs count and vocab size statistics of our training data. PD refers to the People's Daily, TP is TouchPal corpus.

Here is the hyperparameters we used: (a) deep LSTM models, 3 layers, 500 cells, (c) 13 epoch training with plain SGD and a simple learning rate schedule - start with a learning rate of 1.0; after 9 epochs, halve the learning rate every epoch, (d) mini-batches are of size 64 and shuffled, (e) dropout is 0.3. The pre-trained pinyin embeddings and Chinese word embeddings are trained by word2vec (Mikolov et al., 2013) toolkit on Wikipedia[4] and unseen words are assigned unique random vectors.

2.2 IR-based association module

We use IR-based association module to help user type long sentences which can predict the whole expected inputs according to the similarity between user's incomplete input and the candidates in a corpus containing massive sentences. In this work, we use Term Frequency-Inverse Document Frequency (TF-IDF) to calculate the similarity measurement, which has been usually used in text classification and information retrieval. The TF (term-frequency) term is simply a count of the number of times a word appearing in a given context, while the IDF (invert document frequency) term puts a penalty on how often the word appears elsewhere in the corpus. The final TF-IDF score is calculated by the product of these two terms, which is formulated as:

$$\text{TF-IDF}(w, d, D) = f(w, d) \times \log \frac{N}{|\{d \in D : w \in d\}|}$$

where $f(w, d)$ indicates the number of times word w appearing in context d, N is the total number of dialogues, and the denominator represents the number of dialogues in which the word w appears.

In the IME scenario, the TF-IDF vectors are first calculated for the input context and each of the candidate responses from the corpus. Given a set of candidate response vectors, the one with the highest cosine similarity to the context vector is selected as the output. For Recall @ k, the top k candidates are returned. In this work, we only make use of the top 1 matched one.

3 User Experience Advantages

3.1 High Quality of P2C

We utilize Maximum Input Unit (MIU) Accuracy (Zhang et al., 2017) to evaluate the quality of our P2C module by measuring the conversion accuracy of MIU, whose definition is the longest uninterrupted Chinese character sequence inside a sentence. As the P2C conversion aims to output a ranked list of corresponding character sequences candidates, the top-K MIU accuracy means the possibility of hitting the target in the first K predicted items. We will follow the definition of (Zhang et al., 2017) about top-K accuracy.

Our model is compared to other models in Table 2. So far, (Huang et al., 2015) and (Zhang et al., 2017) reported the state-of-the-art results among statistical models. We list the top-5 accuracy contrast to all baselines with top-10 results, and the comparison indicates the noticeable advancement of our P2C model. To our surprise, the top-5 result on PD of our P2C module approaches the top-10 accuracy of Google IME. On DC corpus, the P2C module with the best setting achieves 90.17% accuracy, surpassing all the baselines. The comparison shows the high quality of our P2C conversion.

3.2 Association Cloud Platform

Follow-up Prediction An accurate P2C conversion is only the fundamental requirement to build

[4]https://dumps.wikimedia.org/zhwiki/20180201/zhwiki-20180201-pages-articles-multistream.xml.bz2

	DC			PD		
	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
(Huang et al., 2015)	59.15	71.85	76.78	61.42	73.08	78.33
(Zhang et al., 2017)	57.14	72.32	80.21	64.42	72.91	77.93
Google IME	62.13	72.17	74.72	**70.93**	80.32	**82.23**
P2C of Moon	**71.31**	**89.12**	**90.17**	70.51	**79.83**	80.12

Table 2: Comparison with previous state-of-the-art P2C models.

an intelligent IME which is not only supposed to give accurate P2C conversion, but to help users type sentences in a more convenient and efficient manner. To this end, follow-up prediction is quite necessary for input acceleration. Given an unfinished input, Moon IME now enables the follow-up prediction to help the user complete the typing. For example, given "快速傅里" (Fast Fourier), the IME engine will provide the candidate "快速傅里叶变换" (fast Fourier transform). Specifically, we extract each sentence in the Wikipedia corpus and use the IR-based association module to retrieve the index continuously and give the best-matched sentence as the prediction.

Pinyin-to-English Translation Our Moon IME is also equipped with a multi-lingual typing ability. For users of different language backgrounds, a satisfying conversation can benefit from the direct translation in IME engine. For example, if a Chinese user is using our IME chatting with a native English speaker, but get confused with how to say "Input Method Engine", simply typing the words "输入法" in mother tongue, the IME will give the translated expression. This is also achieved by training a Seq2Seq model from OpenNMT using WMT17 Chinese-English dataset[5].

Factoid Question Answering As an instance of IR-based association module, we make use of question answering (QA) corpus for automatic question completion. Intuitively, if a user wants to raise a question, our IME will retrieve the most matched question in the corpus along with the corresponding answer for typing reference. We use the WebQA dataset (Li et al., 2016) as our QA corpus, which contains more than 42K factoid question-answer pairs. For example, if a user input "吉他有" or "吉他弦" (guitar strings), the candidate "吉他有几根弦" (How many strings are there in the guitar?).

[5]http://www.statmt.org/wmt17/translation-task.html

Figure 2: A case study of Association Cloud Platform.

Figure 2 shows a typical result returned by the platform when a user gives incomplete input. When user input pinyin sequence such as "zui da de ping", the P2C module returns "最大的平" as one candidate of the generated list and sends it to association platform. Then associative prediction is given according to the input mode that user current selections. Since the demands of the users are quite diverse, our platform to support such demands can be adapted to any specific domains with complex specialized terms. We provide a

Demo homepage[6] for better reference, in which we display the main feature function of our platform and provide a download link.

4 Related Work

There are variable referential natural language processing studies(Cai et al., 2018; Li et al., 2018b; He et al., 2018; Li et al., 2018a; Zhang et al., 2018a; Cai et al., 2017a,b) for IME development to refer to. Most of the engineering practice mainly focus on the matching correspondence between the Pinyin and Chinese characters, namely, pinyin-to-character converting with the highest accuracy. (Chen, 2003) introduced a conditional maximum entropy model with syllabification for grapheme-to-phoneme conversion. (Zhang et al., 2006) presented a rule-based error correction approach to improving preferable conversion rate. (Lin and Zhang, 2008) present a statistical model that associates a word with supporting context to offer a better solution to Chinese input. (Jiang et al., 2007) put forward a PTC framework based on support vector machine. (Okuno and Mori, 2012) introduced an ensemble model of word-based and character-based models for Japanese and Chinese IMEs. (Yang et al., 2012; Wang et al., 2018, 2016; Pang et al., 2016; Jia and Zhao, 2013, 2014) regarded the P2C conversion as a transformation between two languages and solved it by statistical machine translation framework. (Chen et al., 2015) firstly use natural machine thanslation method to translate pinyin to Chinese. (Zhang et al., 2017) introduced an online algorithm to construct an appropriate dictionary for IME.

The recent trend on state-of-the-art techniques for Chinese input methods can be put into two lines. Speech-to-text input as iFly IM[7] (Zhang et al., 2015; Saon et al., 2014; Lu et al., 2016) and the aided input methods which are capable of generating candidate sentences for users to choose to complete input tasks, means that users can yield coherent text with fewer keystrokes. The challenge is that the input pinyin sequences are too imperfect to support sufficient training. Most existing commercial input methods offer auto-completion to users as well as extended association functions, to aid users input. However, the performance of association function of existing commercial IMEs are unsatisfactory to relevant user requirement for oversimplified modeling.

It is worth mentioning that we delivery Moon IME as a type of IME service rather than a simple IME software because it can be adjusted to adapt to diverse domains with the Association Cloud Platform (Zhang et al., 2018b,c; Zhang and Zhao, 2018), which helps user type long sentences and predicts the whole expected inputs based on customized knowledge bases.

5 Conclusion

This work makes the first attempt at establishing a general cloud platform to provide customizable association services for Chinese pinyin IME as to our best knowledge. We present Moon IME, a pinyin IME that contains a high-quality P2C module and an extended information retrieval based module. The former is based on an attention-based NMT model and the latter contains follow-up-prediction and machine translation module for typing assistance. With a powerful customizable design, the association cloud platform can be adapted to any specific domains including complex specialized terms. Usability analysis shows that core engine achieves comparable conversion quality with the state-of-the-art research models and the association function is stable and can be well adopted by a broad range of users. It is more convenient for predicting complete, extra and even corrected character outputs especially when user input is incomplete or incorrect.

[6] ime.leisure-x.com

[7] https://www.xunfei.cn/

References

Deng Cai, Hai Zhao, Yang Xin, Yuzhu Wang, and Zhongye Jia. 2017a. A hybrid model for Chinese spelling check. In *ACM Transactions on Asian Low-Resource Language Information Process*.

Deng Cai, Hai Zhao, Zhisong Zhang, Yang Xin, Yongjian Wu, and Feiyue Huang. 2017b. Fast and accurate neural word segmentation for Chinese. In *ACL*, pages 608–615.

Jiaxun Cai, Shexia He, Zuchao Li, and Hai Zhao. 2018. A full end-to-end semantic role labeler, syntax-agnostic or syntax-aware? In *COLING*.

Shenyuan Chen, Rui Wang, and Hai Zhao. 2015. Neural network language model for Chinese pinyin input method engine. In *PACLIC-29*, pages 455–461.

Stanley F. Chen. 2003. Conditional and joint models for grapheme-to-phoneme conversion. In *INTERSPEECH*, pages 2033–2036.

Shexia He, Zuchao Li, Hai Zhao, Hongxiao Bai, and Gongshen Liu. 2018. Syntax for semantic role labeling, to be, or not to be. In *ACL*.

Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Guoping Huang, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2015. A new input method for human translators: integrating machine translation effectively and imperceptibly. In *IJCAI*, pages 1163–1169.

Zhongye Jia and Hai Zhao. 2013. Kyss 1.0: a framework for automatic evaluation of Chinese input method engines. In *IJCNLP*, pages 1195–1201.

Zhongye Jia and Hai Zhao. 2014. A joint graph model for pinyin-to-Chinese conversion with typo correction. In *ACL*, pages 1512–1523.

Wei Jiang, Yi Guan, Xiao Long Wang, and Bing Quan Liu. 2007. Pinyin to character conversion model based on support vector machines. *Journal of Chinese Information Processing*, 21(2):100–105.

Haonan Li, Zhisong Zhang, yuqi Ju, and Hai Zhao. 2018a. Neural character-level dependency parsing for Chinese. In *AAAI*.

Peng Li, Wei Li, Zhengyan He, Xuguang Wang, Ying Cao, Jie Zhou, and Wei Xu. 2016. Dataset and neural recurrent sequence labeling model for open-domain factoid question answering. *arXiv preprint arXiv:1607.06275v2*.

Zuchao Li, Shexia He, and Hai Zhao. 2018b. Seq2seq dependency parsing. In *COLING*.

Bo Lin and Jun Zhang. 2008. A novel statistical Chinese language model and its application in pinyin-to-character conversion. In *ACM Conference on Information and Knowledge Management*, pages 1433–1434.

Liang Lu, Kong Lingpeng, Chris Dyer, Noah A. Smith, and Steve Renals. 2016. Unfolded recurrent neural networks for speech recognition. In *INTERSPEECH*, pages 385–389.

Minh Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *EMNLP*, pages 1412–1421.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Yoh Okuno and Shinsuke Mori. 2012. An ensemble model of word-based and character-based models for Japanese and Chinese input method. In *CoLING*, pages 15–28.

Chenxi Pang, Hai Zhao, and Zhongyi Li. 2016. I can guess what you mean: A monolingual query enhancement for machine translation. In *CCL*, pages 50–63.

George Saon, Hagen Soltau, Ahmad Emami, and Michael Picheny. 2014. Unfolded recurrent neural networks for speech recognition. In *INTERSPEECH*, pages 343–347.

Rui Wang, Hai Zhao, Bao-Liang Lu, Masao Utiyama, and Eiichiro Sumita. 2016. Connecting phrase based statistical machine translation adaptation. In *CoLING*, pages 3135–3145.

Rui Wang, Hai Zhao, Sabine Ploux, Bao-Liang Lu, Masao Utiyama, and Eiichiro Sumita. 2018. Graph-based bilingual word embedding for statistical machine translation. In *ACM Transaciton on Asian and Low-Resource Language Information Processing*, volume 17.

Yu Wu, Wei Wu, Chen Xing, Zhoujun Li, and Ming Zhou. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *ACL*, pages 496–505.

Shaohua Yang, Hai Zhao, and Bao-liang Lu. 2012. A machine translation approach for Chinese whole-sentence pinyin-to-character conversion. In *PACLIC-26*, pages 333–342.

Shiliang Zhang, Cong Liu, Hui Jiang, Si Wei, Lirong Dai, and Yu Hu. 2015. Feedforward sequential memory networks: A new structure to learn long-term dependency. *arXiv preprint arXiv:1512.08301*.

Xihu Zhang, Chu Wei, and Hai Zhao. 2017. Tracing a loose wordhood for Chinese input method engine. *arXiv preprint arXiv:1712.04158*.

Yan Zhang, Bo Xu, and Chengqing Zong. 2006. Rule-based post-processing of pin to Chinese characters conversion system. In *International Symposium on Chinese Spoken Language Processing*, pages 1–4.

Zhuosheng Zhang, Jiangtong Li, Hai Zhao, and Bingjie Tang. 2018a. Sjtu-nlp at semeval-2018 task 9: Neural hypernym discovery with term embeddings. In *SemEval-2018, Workshop of NAACL-HLT*.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, and Hai Zhao. 2018b. Modeling multi-turn conversation with deep utterance aggregation. In *CoLING*.

Zhuosheng Zhang, Huang Yafang, and Hai Zhao. 2018c. Subword-augmented embedding for cloze reading comprehension. In *CoLING*.

Zhuosheng Zhang and Hai Zhao. 2018. One-shot learning for question-answering in gaokao history challenge. In *CoLING*.

Hai Zhao, Chang-Ning Huang, Mu Li, and Taku Kudo. 2006. An improved Chinese word segmentation system with conditional random field. In *Sighan*, pages 162–165.