

# Chinese Pinyin Aided IME, Input What You Have Not Keystroked Yet

Yafang Huang<sup>1,2</sup>, Hai Zhao<sup>1,2,\*</sup>,

<sup>1</sup>Department of Computer Science and Engineering,  
Shanghai Jiao Tong University, Shanghai, 200240, China

<sup>2</sup>Key Laboratory of Shanghai Education Commission for Intelligent Interaction  
and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China  
huangyafang@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn \*

## Abstract

Chinese pinyin input method engine (IME) converts pinyin into character so that Chinese characters can be conveniently inputted into computer through common keyboard. IMEs work relying on its core component, pinyin-to-character conversion (P2C). Usually Chinese IMEs simply predict a list of character sequences for user choice only according to user pinyin input at each turn. However, Chinese inputting is a multi-turn online procedure, which can be supposed to be exploited for further user experience promoting. This paper thus for the first time introduces a sequence-to-sequence model with gated-attention mechanism for the core task in IMEs. The proposed neural P2C model is learned by encoding previous input utterance as extra context to enable our IME capable of predicting character sequence with incomplete pinyin input. Our model is evaluated in different benchmark datasets showing great user experience improvement compared to traditional models, which demonstrates the first engineering practice of building Chinese aided IME.

## 1 Introduction

Pinyin is the official romanization representation for Chinese and the P2C converting the inputted pinyin sequence to Chinese character sequence is the most basic module of all pinyin based IMEs.

Most of the previous research (Chen, 2003; Zhang et al., 2006; Lin and Zhang, 2008; Chen and Lee, 2000; Jiang et al., 2007; Cai et al., 2017a) for IME focused on the matching correspondence

between pinyin syllables and Chinese characters. (Huang et al., 2018; Yang et al., 2012; Jia and Zhao, 2014; Chen et al., 2015) regarded the P2C as a translation between two languages and solved it in statistical or neural machine translation framework. The fundamental difference between (Chen et al., 2015) work and ours is that our work is a fully end-to-end neural IME model with extra attention enhancement, while the former still works on traditional IME only with converted neural network language model enhancement. (Zhang et al., 2017) introduced an online algorithm to construct appropriate dictionary for P2C. All the above mentioned work, however, still rely on a complete input pattern, and IME users have to input very long pinyin sequence to guarantee the accuracy of P2C module as longer pinyin sequence may receive less decoding ambiguity.

The Chinese IME is supposed to let user input Chinese characters with least inputting cost, i.e., keystroking, which indicates extra content predication from incomplete inputting will be extremely welcomed by all IME users. (Huang et al., 2015) partially realized such an extra predication using a maximum suffix matching postprocessing in vocabulary after SMT based P2C to predict longer words than the inputted pinyin.

To facilitate the most convenience for such an IME, in terms of a sequence to sequence model as neural machine translation (NMT) between pinyin sequence and character sequence, we propose a P2C model with the entire previous inputted utterance confirmed by IME users being used as a part of the source input. When learning the type of the previous utterance varies from the previous sentence in the same article to the previous turn of utterance in a conversation, the resulting IME will make amazing predication far more than what the pinyin IME users actually input.

In this paper, we adopt the attention-based NMT

\* 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), and the joint research project with YouTu Lab of Tencent.

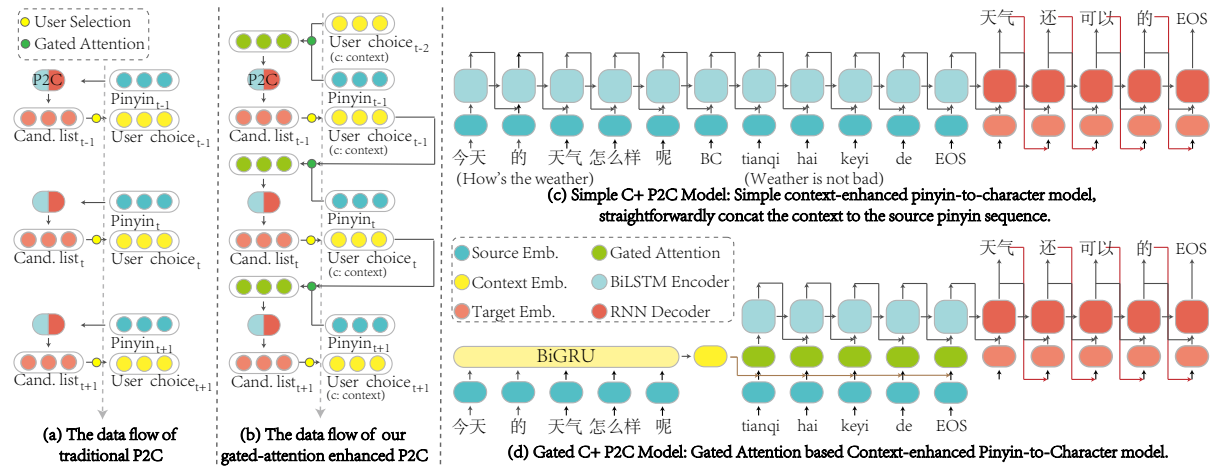


Figure 1: Architecture of the proposed model.

framework in (Luong et al., 2015) for the P2C task. In contrast to related work that simply extended the source side with different sized context window to improve of translation quality (Tiedemann and Scherrer, 2017), we add the entire input utterance according to IME user choice at previous time (referred to the *context* hereafter). Hence the resulting IME may effectively improve P2C quality with the help of extra information offered by context and support incomplete pinyin input but predict complete, extra, and corrected character output. The evaluation and analysis will be performed on two Chinese datasets, include a Chinese open domain conversations dataset for verifying the effectiveness of the proposed method.

## 2 Model

As illustrated in Figure 1, the core of our P2C is based the attention-based neural machine translation model that converts at word level. Still, we formulize P2C as a translation between pinyin and character sequences as shown in a traditional model in Figure 1(a). However, there comes a key difference from any previous work that our source language side includes two types of inputs, the current source pinyin sequence (noted as  $P$ ) as usual, and the extended context, i.e., target character sequence inputted by IME user last time (noted as  $C$ ). As IME works dynamically, every time IME makes a predication according to a source pinyin input, user has to indicate the 'right answer' to output target character sequence for P2C model learning. This online work mode of IMEs can be fully exploited by our model whose work flow is shown in Figure 1(b).

As introduced a hybrid source side input, our model has to handle document-wide translation by considering discourse relationship between two consecutive sentences. The most straightforward modeling is to simply concatenate two types of source inputs with a special token 'BC' as separator. Such a model is in Figure 1(c). However, the significant drawback of the model is that there are a slew of unnecessary words in the extended context (previous utterance) playing a noisy role in the source side representation.

To alleviate the noise issue introduced by the extra part in the source side, inspired by the work of (Dhingra et al., 2016; Pang et al., 2016; Zhang et al., 2018c,a,b; Cai et al., 2017b), our model adopts a gated-attention (GA) mechanism that performs multiple hops over the pinyin with the extended context as shown in Figure 1(d). In order to ensure the correlation between each other, we build a parallel bilingual training corpus and use it to train the pinyin embeddings and the Chinese embeddings at once. We use two Bidirectional gated recurrent unit (BiGRU) (Cho et al., 2014) to get contextual representations of the source pinyin and context respectively,  $H_p = \text{BiGRU}(P)$ ,  $H_c = \text{BiGRU}(C)$ , where the representation of each word is formed by concatenating the forward and backward hidden states.

For each pinyin  $p_i$  in  $H_p$ , the GA module forms a word-specific representation of the context  $c_i \in H_c$  using soft attention, and then adopts element-wise product to multiply the context representation with the pinyin representation.  $\alpha_i = \text{softmax}(H_c^T p_i)$ ,  $\beta_i = C \alpha_i$ ,  $x_i = p_i \odot \beta_i$ , where  $\odot$  is multiplication operator.

The pinyin representation  $\tilde{H}_p = x_1, x_2, \dots, x_k$

is augmented by context representation and then sent into the encoder-decoder framework. The encoder is a bi-directional long short-term memory (LSTM) network (Hochreiter and Schmidhuber, 1997). The vectorized inputs are fed to forward and backward LSTMs to obtain the internal representation of two directions. The output for each input is the concatenation of the two vectors from both directions. Our decoder based on the global attentional models proposed by (Luong et al., 2015) to consider the hidden states of the encoder when deriving the context vector. The probability is conditioned on a distinct context vector for each target word. The context vector is computed as a weighted sum of previous 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 embeddings and candidate target characters.

This work belongs to one of the first line which fully introduces end-to-end deep learning solution to the IME implementation following a series of our previous work (Zhu et al., 2018; Wu and Zhao, 2018; Zhang and Zhao, 2018; Bai and Zhao, 2018; Cai et al., 2018; He et al., 2018; Qin et al., 2018; Li et al., 2018; Jia et al., 2013).

### 3 Experiment

#### 3.1 Definition of Incomplete Input for IME

The completeness in IME is actually uneasily well-defined as it is a relative concept for inputting procedure. Note that not until user types the return key *enter*, user will not (usually) really make the input choice. Meanwhile, even though the entire/complete input can be strictly defined by the time when user types *enter*, user still can make decision at any time and such incompleteness cannot be well evaluated by all the current IME metrics. As the incomplete from is hard to simulate and it is diverse in types, we have to partially evaluate it in the following two ways<sup>1</sup>,

#### The incomplete pinyin as abbreviation pinyin

To compare with previous work directly, we followed (Huang et al., 2015) and focused on the abbreviated pinyin (the consonant letter only) to perform evaluation (i.e., tian.qi to t.q).

#### Take incomplete user input as the incomplete

As IME works as an interactive system, it will always give prediction only if users keep typing. If

user’s input does not end with typing *enter*, we can regard the current input pinyin sequence is an incomplete one.

### 3.2 Datasets and Settings

	PD		DC	
	Train	Test	Train	Test
# Sentence	5.0M	2.0K	1.0M	2.0K
L < 10 %	88.7	89.5	43.0	54.0
L < 50 %	11.3	10.5	47.0	42.0
L > 50 %	0.0	0.0	4.0	2.0
Relativity %	18.0	21.1	65.8	53.4

Table 1: Data statistics for the sentence number and sentence length (L) of two corpora.

Our model is evaluated on two datasets, namely the People’s Daily (PD) corpus and Douban conversation (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. The relativity refers to total proportion of sentences that associate with contextual history at word level. For example, there are 65.8% of sentences of DC corpus have words appearing in the context. 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).

Our model was implemented using the PyTorch<sup>2</sup> library, here is the hyperparameters we used: (a) the RNNs used are 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. Word embeddings are pre-trained by word2vec (Mikolov et al., 2013) toolkit on the adopted corpus and unseen words are assigned unique random vectors. (f) the gated attention layers size is 3, the hidden units number of BiGRU is 100.

Two metrics are used: Maximum Input Unit (MIU) accuracy (Zhang et al., 2017) and KeyStroke Score (KySS) (Jia and Zhao, 2013).

<sup>1</sup>Our code is at <https://github.com/YvonneHuang/galME>

<sup>2</sup><https://github.com/pytorch/pytorch>

	DC		PD	
	Top-1	KySS	Top-1	KySS
CoCat★	49.72	0.7393	48.78	0.7868
Basic P2C★	52.31	0.7305	47.95	0.7879
Simple C+ P2C★	23.83	0.5431	43.95	0.7495
Gated C+ P2C★	53.76	0.8896	48.31	0.7931
CoCat	59.15	0.7651	61.42	0.7933
Basic P2C	71.31	0.8845	70.5	0.8301
Simple C+ P2C	61.28	0.7963	60.87	0.7883
Gated C+ P2C	<b>73.89</b>	<b>0.8935</b>	<b>70.98</b>	<b>0.8407</b>

Table 2: The effect of P2C modules with Different Input Forms. ★ means that the input is incomplete.

The former measures the conversion accuracy of MIU, whose definition is the longest uninterrupted Chinese character sequence during inputting. As the P2C conversion aims to output a rank list of the corresponding character sequences candidates, the Top- $K$  MIU accuracy means the possibility of hitting the target in the first  $K$  predicted items. The KySS quantifies user experience by using keystroke count. For an ideal IME with complete input, we have  $KySS = 1$ . An IME with higher KySS is supposed to perform better.

### 3.3 Model Definition

We considered the following baselines: (a) Google IME: the only commercial Chinese IME providing a debuggable API in the market now; (b) OMWA: online model for word acquisition proposed by (Zhang et al., 2017); (c) CoCat: an SMT based input method proposed by (Huang et al., 2015) that supports incomplete pinyin inputs.

Three models with incomplete or complete inputs will be evaluated: (a) Basic P2C, the basic P2C based on attention-NMT model; (b) Basic C2C, the basic C to C model based on Seq2Seq model; (b) Simple C+ P2C, the simple concatenated P2C conversion model that concatenate context to pinyin representation; (c) Gated C+ P2C, our gated attention based context-enhanced pinyin-to-character model. Pinyin in model \* has been actually set to abbreviation form when we say it goes to (Huang et al., 2015) incomplete definition.

### 3.4 Result and Analysis

**Effect of Gated Attention Mechanism** Table 3 shows the Effect of gated attention mechanism. We compared models with Gated C+ P2C and Simple C+ P2C. The MIU accuracy of the P2C

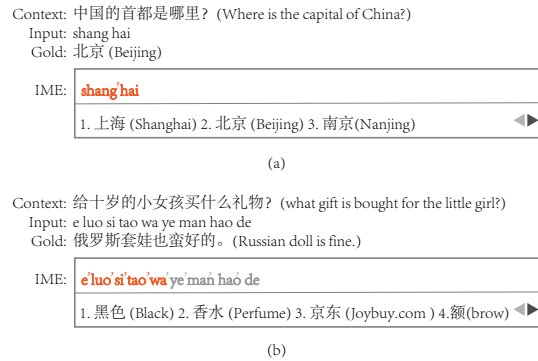


Figure 2: Examples of the candidates list given by the proposed IMEs.

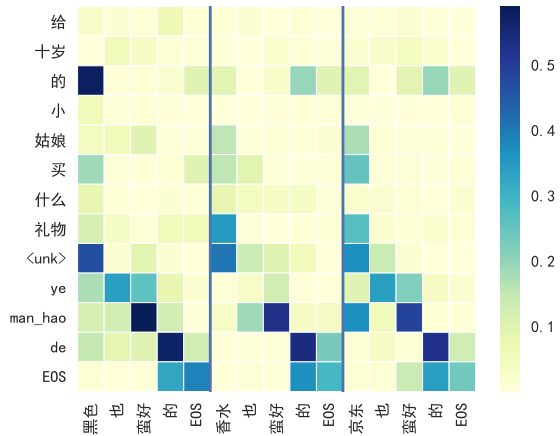


Figure 3: Attention visualization. Deeper color mean larger value.

model has over 10% improvement when changing the operate pattern of the extra information proves the effect of GA mechanism. The Gated C+ P2C achieves the best in DC corpus, suggesting that the gated-attention works extremely well for handling long and diverse context.

**Effect of P2C modules with Different Input Forms** Table 2 shows the evaluation results of P2C modules with different input forms. It should

	DC				PD			
	Top-1	Top-5	Top-10	KySS	Top-1	Top-5	Top-10	KySS
Google IME	62.13	72.17	74.72	0.6731	70.93	80.32	<b>82.23</b>	0.7535
CoCat	59.15	71.85	76.78	0.7651	61.42	73.08	78.33	0.7933
OMWA	57.14	72.32	80.21	0.7389	64.42	72.91	77.93	0.7115
basic P2C	71.31	89.12	90.17	0.8845	70.5	79.8	80.1	0.8301
Simple C+ P2C	61.28	71.88	73.74	0.7963	60.87	71.23	75.33	0.7883
Gated C+ P2C	<b>73.89</b>	<b>90.14</b>	<b>91.22</b>	<b>0.8935</b>	<b>70.98</b>	<b>80.79</b>	81.37	<b>0.8407</b>

Table 3: Comparison with previous state-of-the-art models.

not surprise that straightforward concatenation strategy for source inputs performs poorly when the input pinyin is incomplete in DC corpus, due to obvious noise in too long context. The relatively small gap between the results of CoCat\* and CoCat indicate that statistical learning model may be helpful in obtaining some useful patterns from limited input. When the input statement contains adequacy information, the MIU accuracy of Gated C+ P2C system achieves more than 20% improvement in both corpora. However, we find that the KySS scores are much more close even with different pinyin integrity, which indicates that user experience in terms of KySS are more hard improved.

**Instance Analysis** We input a dialogue in wonder to how much of the contextual information is used when P2C module find the input pinyin is unknown. Figure 2 demonstrates the effect of the gated attention mechanism on candidates offering and unknown word replacement. As shown in Figure 2(a), we find that our IME suggests a more suitable candidates to the user when user is obviously not consistent with what the model has learned previously, which shows our model exceeds the Simple C+ P2C learning for maximally matching the inputted pinyin, but become capable of effectively resisting user pinyin input noise, and turns to learn potential language knowledge in previous input history<sup>3</sup>.

As the ability predict user input from incomplete pinyin cannot be covered by any current IME performance metrics, thus the reported results yielded by our model actually underestimate our model performance to some extent. We illustrate the empirical discoveries of Figure 2(b) to

<sup>3</sup>Note as we evaluate our model only on two available corpora, but not the real world case from true user inputting history, which makes the instance situation limit to the domain feature of the given corpora.

demonstrate the extra effect of our P2C system on such situation, which indicates that the gated-attention pattern has taken great advantage of contextual information when given an unknown word. Or, namely, our model enables the incomplete input prediction though has to let it outside the current IME performance measurement. We display the attention visualization of Figure 2(b) in Figure 3 for better reference to explain the effect extended context plays on the generation of target characters.

**Main Result** Our model is compared to other models in Table 3. 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 best Gated C+ P2C system approaches the top-10 accuracy of Google IME. On DC corpus, our Gated C+ P2C model with the best setting achieves 90.14% accuracy, surpassing all the baselines. The comparison shows our gated-attention system outperforms all state-of-the-art baselines with better user experience.

## 4 Conclusion

For the first time, this work makes an attempt to introduce additional context in neural pinyin-to-character converter for pinyin-based Chinese IME as to our best knowledge. We propose a gated-attention enhanced model for digging significant context information to improve conversion quality. More importantly, the resulting IME supports incomplete user pinyin input but returns complete, extra and even corrected character outputs, which brings about a story-telling mode change for all existing IMEs.

## References

- Hongxiao Bai and Hai Zhao. 2018. Deep enhanced representation for implicit discourse relation recognition. In *CoLING*, pages 571–583.
- Deng Cai, Hai Zhao, xin Yang, Wang Yuzhu, and Jia Zhongye. 2017a. A hybrid model for Chinese spelling check. In *ACM Transactions on Asian Low-Resource Language Information Process*.
- Deng Cai, Hai Zhao, Zhisong Zhang, Yuan Xin, Yongjiang 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*, pages 2753–2765.
- 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 *Eurospeech 2003 - INTERSPEECH 2003*, pages 2033–2036.
- Zheng Chen and Kai Fu Lee. 2000. A new statistical approach to Chinese pinyin input. In *ACL*, pages 241–247.
- Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, pages 1724–1734.
- Bhuvan Dhingra, Hanxiao Liu, Zhilin Yang, William W Cohen, and Ruslan Salakhutdinov. 2016. Gated-attention readers for text comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1832–1846.
- 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*, pages 2061–2071.
- 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.
- Yafang Huang, Zuchao Li, Zhuosheng Zhang, and Hai Zhao. 2018. Moon ime: Neural-based Chinese pinyin aided input method with customizable association. In *ACL: System Demonstrations*, pages 140–145.
- Zhongye Jia, Peilu Wang, and Hai Zhao. 2013. Graph model for Chinese spell checking. In *SIGHAN-7*, pages 88–92.
- 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.
- Zuchao Li, Jiaxun Cai, Shexia He, and Hai Zhao. 2018. Seq2seq dependency parsing. In *CoLING*, pages 3203–3214.
- 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.
- 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*.
- Chenxi Pang, Hai Zhao, and Zhongyi Li. 2016. I can guess what you mean: A monolingual query enhancement for machine translation. In *CCL*, volume 10035, pages 50–63.
- Lianhui Qin, Lemao Liu, Victoria Bi, Yan Wang, Xiaojiang Liu, Zhiting Hu, Hai Zhao, and Shuming Shi. 2018. Automatic article commenting: the task and dataset. In *ACL*, pages 151–156.
- Jorg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92.
- Yingting Wu and Hai Zhao. 2018. Finding better subword segmentation for neural machine translation. In *CCL*.
- Yu Wu, Wei Wu, Chen Xing, Zhoujun Li, and Ming Zhou. 2017. Lsequential 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 *Proceedings of the 26th Asian Pacific conference on language and information and computation*, pages 333–342.

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

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, Yafang Huang, and Hai Zhao. 2018a. Subword-augmented embedding for cloze reading comprehension. In *CoLING*, pages 3740–3752.

Zhuosheng Zhang, Yafang Huang, Pengfei Zhu, and Hai Zhao. 2018b. Effective character-augmented word embedding for machine reading comprehension. In *NLPCC*.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, and Hai Zhao. 2018c. Modeling multi-turn conversation with deep utterance aggregation. In *CoLING*, pages 1802–1814.

Zhuosheng Zhang and Hai Zhao. 2018. One-shot learning for question-answering in gaokao history challenge. In *CoLING*, pages 449–4612.

Pengfei Zhu, Zhuosheng Zhang, Jiangtong Li, Yafang Huang, and Hai Zhao. 2018. Lingke: A fine-grained multi-turn chatbot for customer service. In *CoLING: System Demonstrations*, pages 108–112.

## Appendices

### A Architecture of the Evaluation Models

