Yuankang SUN, Bing LI, Lexiang LI, Peng YANG, Dongmei YANG, 2025. Shared-weight multimodal translation model for recognizing Chinese variant characters. *Frontiers of Information Technology & Electronic Engineering*, 26(7):1066-1082. https://doi.org/10.1631/FITEE.2400504

Shared-weight multimodal translation model for recognizing Chinese variant characters

Key words: Chinese variant characters; Multimodal model; Translation model; Phonology and morphology

Corresponding author: Peng YANG
E-mail: pengyang@seu.edu.cn
ORCID: https://orcid.org/0000-0002-1184-8117

Motivation

The task of recognizing Chinese variant characters aims to address the challenges of semantic ambiguity and confusion, which potentially cause risks to the security of Web content and complicate the governance of sensitive words. We propose a shared-weight multimodal translation model (SMTM) based on multimodal information of Chinese characters, which integrates the phonology of Pinyin and the morphology of fonts into each Chinese character token to learn the deeper semantics of variant texts.

Main idea

- To address the challenges of semantic ambiguity and confusion inherent in Chinese variant characters, we integrate character-level phonological and morphological features using bidirectional encoder representations from Transformers (BERT) embedding, thereby enhancing the efficacy of unimodal solutions.
- We propose a shared-weight embedding mechanism to initialize decoder and generator weights based on the multimodal weights of the source text, strengthening the connection between source and target characters, thereby facilitating target sentence generation.
- We perform simulations using the variant dataset, yielding extensive results that demonstrate the attainment of several state-of-the-art performances by our method.

Method

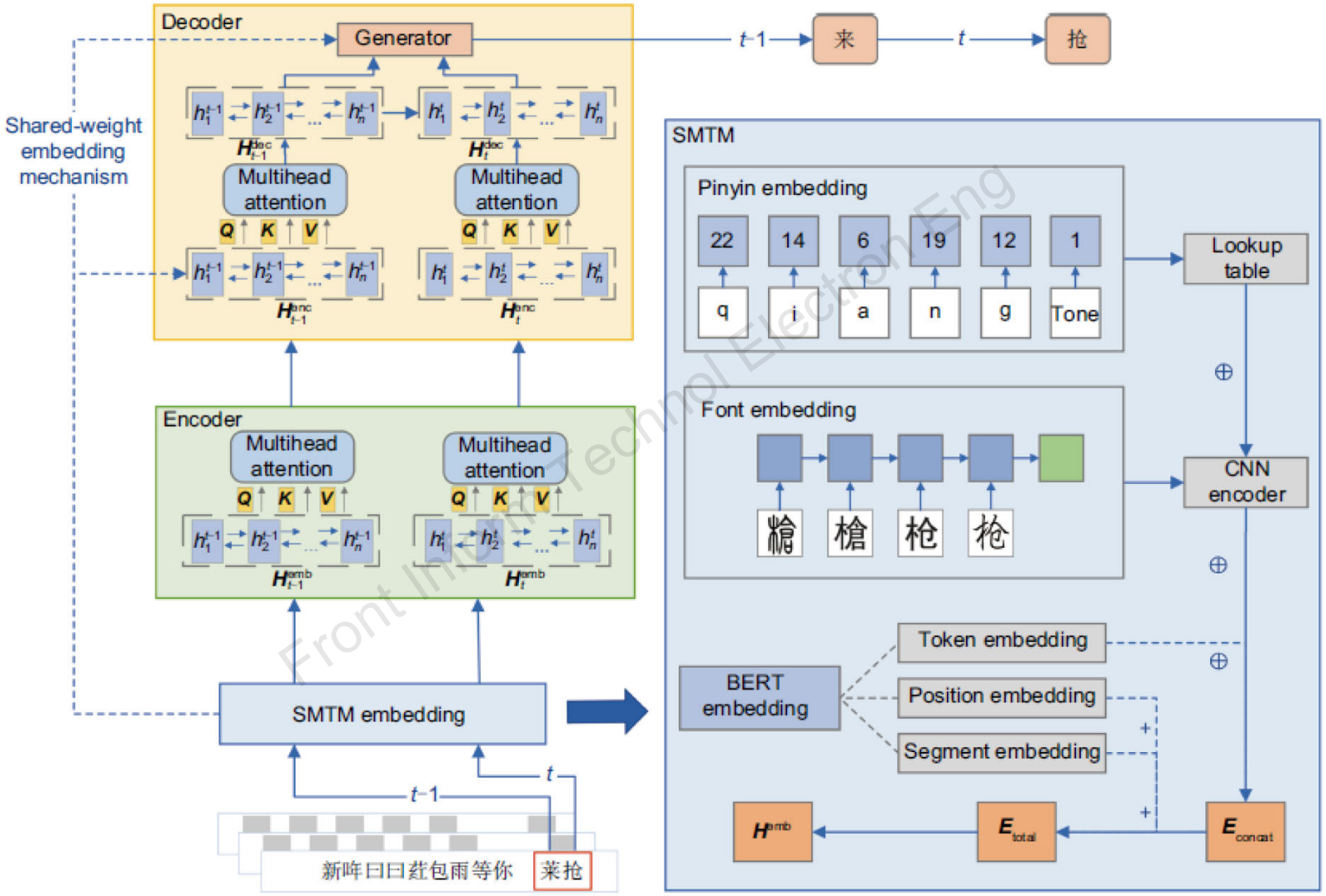

Fig. 1 Illustration of the SMTM architecture, where ⊕ denotes vector concatenation

Method

Algorithm 1 Shared-weight multimodal embedding module

Require: source sentence of Chinese variant characters, **sentence**$=(c_1, c_2, \ldots, c_l)$

Ensure: the vector after the extraction of phonological and morphological features, c_i^{total}

for $i = 0$ **to** len(**sentence**) **do**
  $c_i^{\text{pinyin}} = \text{pinyinProcess}(c_i)$;
  while $\text{len}(c_i^{\text{pinyin}}) \leq 8$ **do**
    $c_i^{\text{pinyin}}.\text{append}(0)$;
  end while
  for $j = 0$ **to** 4 **do**
    $c_j^i = \text{CNN}(c_i)$;
    $c_i^{\text{font}}.\text{append}(c_j^i)$;
  end for
  $(c_i^{\text{token}}, c_i^{\text{pos}}, c_i^{\text{seg}}) \Leftarrow \text{BERTEmbProcess}(c_i)$;
  $c_i^{\text{concat}} = \text{Concat}[c_i^{\text{pinyin}}, c_i^{\text{font}}, c_i^{\text{token}}]$; /* concatenated character representation */
  $c_i^{\text{total}} = c_i^{\text{concat}} + c_i^{\text{pos}} + c_i^{\text{seg}}$;
end for
return c_i^{total}

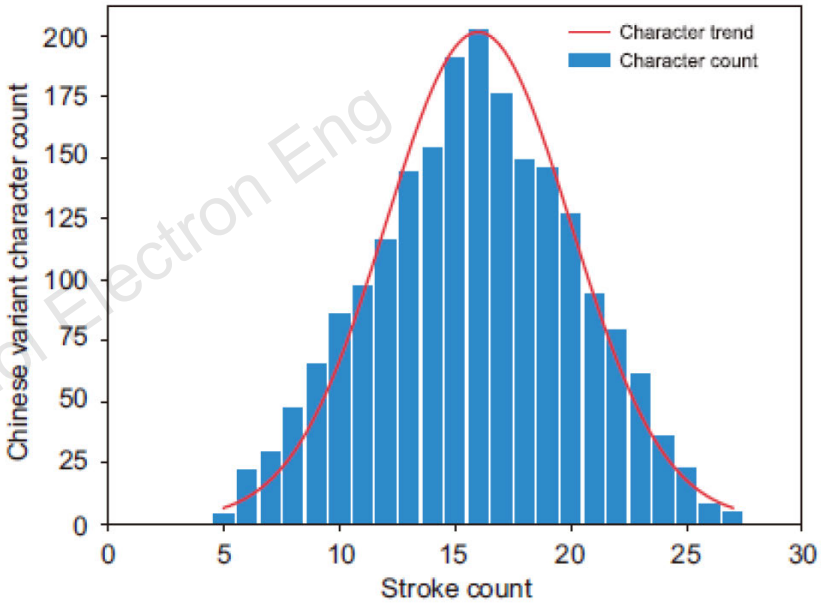

Fig. 2 Distribution of the stroke count and the Chinese variant character count

Results

Table 3 Performance comparison on the CCF BDCI dataset

Type	Model	$BLEU_{1-2}$ (%)	$BLEU_{avg}$ (%)	F1 (%)	Score (%)
Error correction model	CPN	79.767	75.849	70.134	73.971
Seq2Seq	Recurrent Seq2Seq	70.690	65.886	65.510	66.899
	Convolutional Seq2Seq	72.269	68.533	66.667	68.534
	BBT	87.333	83.150	67.027	76.134
Embedding layer	RTT	85.206	80.745	76.392	79.684
	BTT	85.121	80.832	79.006	80.991
Large language model	Llama 3.1 8b	72.871	67.609	65.928	68.084
	Qwen2-7B	81.537	77.132	75.264	77.299
Our model	SMTM	**89.550**	**86.017**	**79.480**	**83.632**

The best results are in bold

Results

Table 4 Results of SMTM model ablation simulations

Model	$BLEU_{1-2}$ (%)	$BLEU_{avg}$ (%)	F1 (%)	Score (%)
SMTM	**89.550**	**86.017**	79.480	**83.632**
W/o Pinyin embedding	88.144 (↓1.406)	84.314 (↓1.703)	75.719 (↓3.761)	80.974 (↓2.658)
W/o font embedding	87.009 (↓2.541)	83.164 (↓2.853)	**81.859** (↑2.379)	83.473 (↓0.159)
W/o sharing for decoder	87.167 (↓2.383)	82.467 (↓3.550)	80.092 (↑0.612)	82.455 (↓1.177)
W/o sharing for generator	87.245 (↓2.305)	82.732 (↓3.285)	79.976 (↑0.496)	82.482 (↓1.150)
W/o emoji enhancement	87.663 (↓1.887)	83.455 (↓2.562)	80.299 (↑0.819)	82.929 (↓0.703)

The best results are in bold. The value in the brackets means the improvement or decrease compared with the SMTM, denoted by ↑ and ↓, respectively

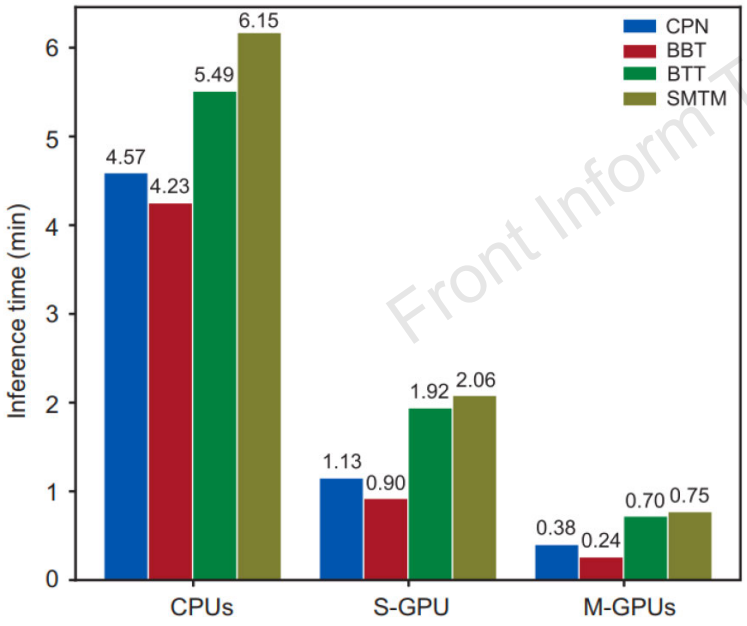

Fig. 3 Comparison of the results of model's efficiency

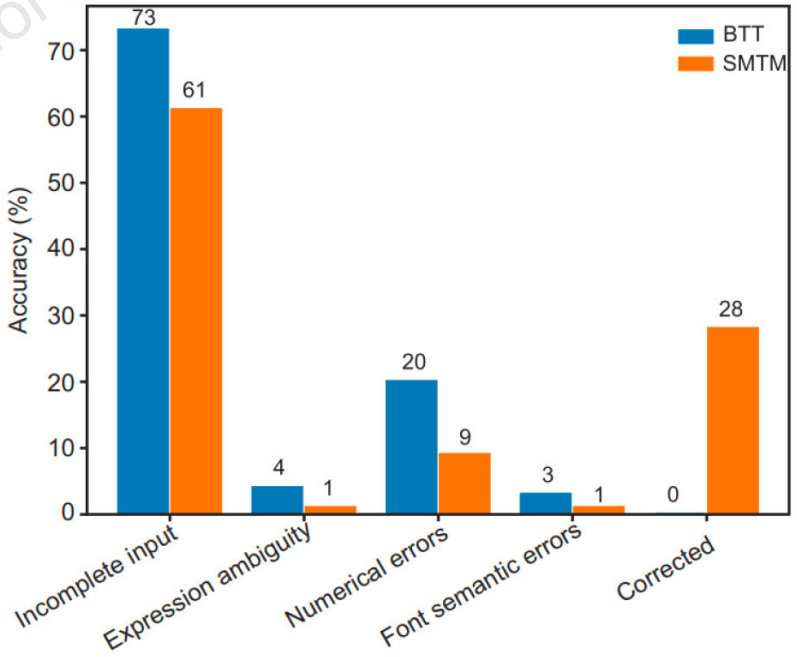

Fig. 4 Error analysis results with baseline BTT as the standard

Results

Variant mode	Text	BTT	SMTM
Homophone	刷单加威信	刷单加微信	刷单加微信
Pinyin substitution	刷单加wei信	刷单加微信	
Pinyin abbreviation	刷单加wx	刷单加wx	
Traditional character substitution	刷單加微信	刷unk加微信	
Similar character substitution	刷单加徽信	刷单加灰心	
Radical split	刷单力口微信	刷单利口微信	
Special symbol substitution	刷单+微信	刷单+微信	
Chinese–English mixture	刷单+V	刷单+V	
Inversion of word order	刷单加信微	刷单加欣慰	

Fig. 5 Case study

Conclusions

In this work, we present the SMTM. It mines the deep semantic features of Chinese variant texts based on Chinese characters, Pinyin, and font images, and uses a shared-weight embedding mechanism to generate target sentences. In the simulations, we show that our model outperforms all compared models in the Chinese variant-character-recognition task. In the investigation of the Chinese variant-character recognition task, precise classification and labeling of variant and standard characters are particularly crucial.

Yuankang SUN obtained his M.S. degree from The University of Sheffield, United Kingdom, in 2021. He is currently pursuing a Ph.D. degree at the School of Computer Science and Engineering, Southeast University and Key Laboratory of Computer Network and Information Integration, Ministry of Education, China, encompassing research focus on artificial intelligence, natural language processing, image processing, and associated domains.

Bing LI received the Ph.D. degree from Southeast University, China, in 2024. She works as a lecturer in the School of Information Technology and Artificial Intelligence at Zhejiang University of Finance and Economics. Her research interests include artificial intelligence, natural language processing, and image processing.

Lexiang LI graduated from Southeast University, China, in 2023. His research directions include natural language processing and Chinese variant character recognition.

PENG YANG received the Ph.D. degree from Southeast University, in 2006, by taking a successive postgraduate and doctoral program. He worked as a Research Scientist with CERN to participate in the Alpha Magnetic Spectrometer (AMS) experiment (PI: Nobel laureate Samuel C.C. Ting), from 2007 to 2009. He is currently a Professor with the School of Computer Science and Engineering, Southeast University, where he is also a Deputy Director of the Future Network Research Center. He is a member of the National Technical Committee of Standardization Administration of China. His research interests focus on new-generation network architecture, edge computing, natural language processing, blockchain, and cyber content governance.

Dongmei YANG is currently with the Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, China. She is also a member of the System Certification Committee of China Cybersecurity Review, Certification and Market Regulation Big Data Center. Her research interests include cybersecurity, satellite and wireless communication, IoT, innovative identification, and global traceability technologies.