

**LREC - COLING 2024**

**Torino, Italy  
20-25 May 2024**



# **An Unsupervised Framework for Adaptive Context-aware Simplified-Traditional Chinese Character Conversion**

**Wei Li<sup>†</sup>, Shutan Huang<sup>†</sup>, Yanqiu Shao**

School of Information Science, Beijing Language and Culture University  
Xueyuan Road 15th, Haidian District, Beijing, China  
liweitj47@blcu.edu.cn, shutan2022@163.com, shaoyanqiu@blcu.edu.cn

---

## Catalogue

0、 Introduction

1、 Approach

2、 Experiment

- Baseline
- Ablation Study
- Analysis
- Case Study

3、 Conclusion



0

# Introduction

## Introduction

Simplified Chinese: “学而时习之，不亦说乎？有朋自远方来，不亦乐乎？人不知而不愠，不亦君子乎？”

Traditional Chinese: 「學而時習之，不亦說乎？有朋自遠方來，不亦樂乎？人不知而不愠，不亦君子乎？」

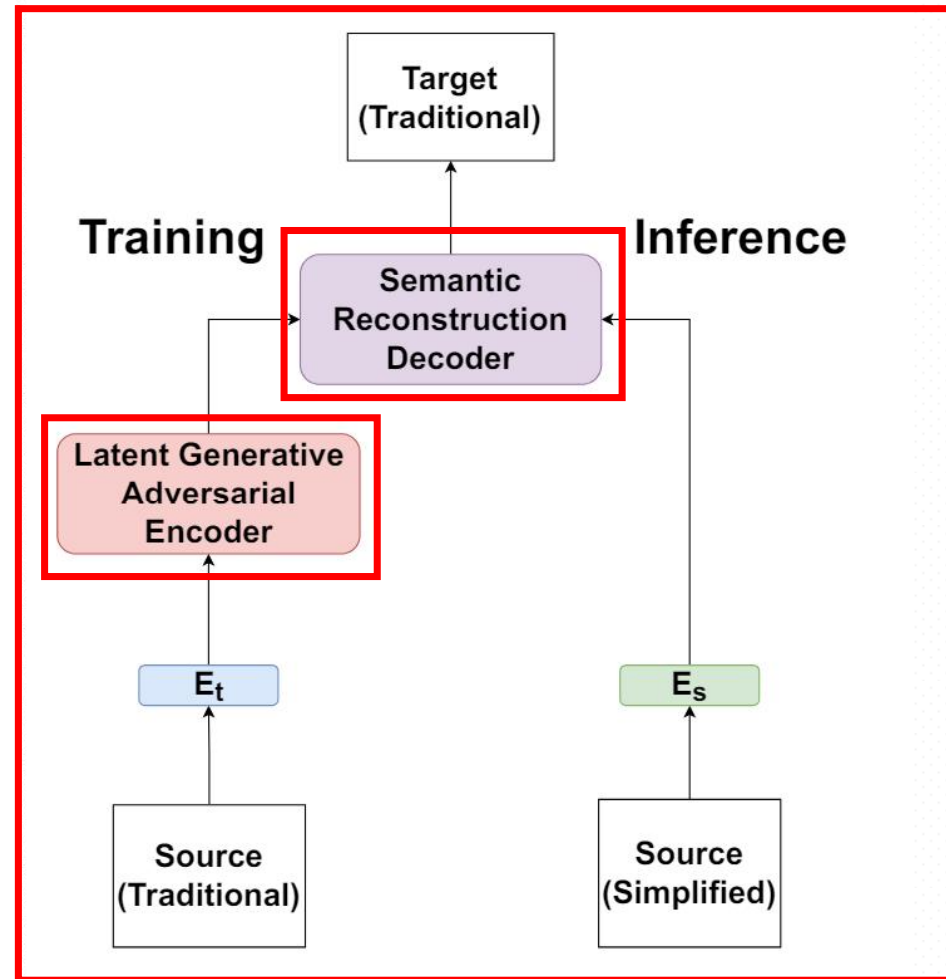
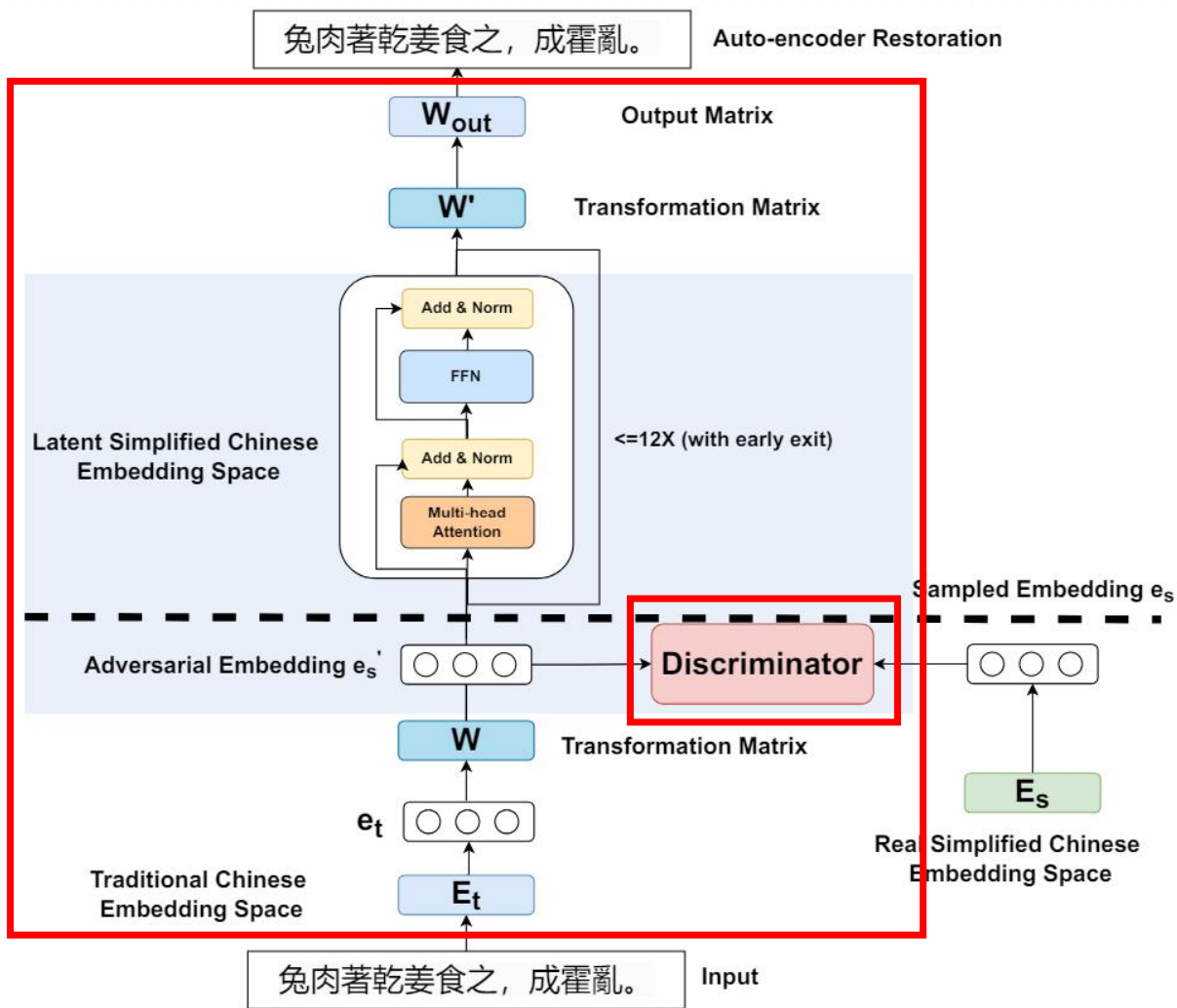
Simplified	Traditional
<p>故国有贤君，折冲万里。</p> <p>中外若一，事无表里。</p> <p>文貌情用，相为内外表里。</p> <p>表里相资，古今一也。</p>	<p>故國有賢君，折沖萬里。</p> <p>中外若一，事無表裏。</p> <p>文貌情用，相為內外表裡。</p> <p>表里相資，古今一也。</p>
<p>星莫大于大辰，北斗常星。</p> <p>朝有变色之言，则下有争斗之患。</p>	<p>星莫大於大辰，北斗常星。</p> <p>朝有變色之言，則下有爭鬥之患。</p>
<p>百官饥饿，河内太守张杨使数千人负米贡饷。</p> <p>四时不出，天下大饥。</p>	<p>百官飢餓，河內太守張楊使數千人負米貢餉。</p> <p>四時不出，天下大饑。</p>

- One-to-many conversion problems
- Large amount of labeled training data
- OOV problem
- Training data cover a long period of time



1

# Approach



2

# Experiment

## Experiment

Method	S-to-T	T-to-S
OpenCC	95.64	97.56
MS Word	96.89	97.51
zhconv	97.96	99.49
pylangtools	96.77	97.79
Website		
Jianfan	94.26	94.98
AIES	97.72	98.84
KJSON	95.72	97.24
Unsupervised MT		
SVD	89.61	89.82
MUSE (Lample et al., 2018)	93.91	93.93
Artetxe et al. (2018a)	93.74	93.88
Artetxe et al. (2018b)(unsup)	94.22	94.21
Artetxe et al. (2018b)(semi)	94.25	94.28
ChatGPT	79.22	79.21
<b>Proposal(base)</b>	<b>98.45</b>	<b>99.51</b>
<b>Proposal(large)</b>	<b>98.57</b>	<b>99.61</b>

## Ablation Study

Module	S-to-T	T-to-S
Latent Generative Adversarial Encoder	93.91	93.93
+ Context-aware Semantic Reconstruction Decoder	96.79	97.66
+ Encoder-Decoder Parameter Sharing	97.13	98.01
+ Early Exit	98.45	99.52
+ OOV Skipping	98.57	99.61

Table 2: This table displays the results of an ablation study where modules were added one by one to evaluate their impact on the performance of the proposed model.

## Analysis

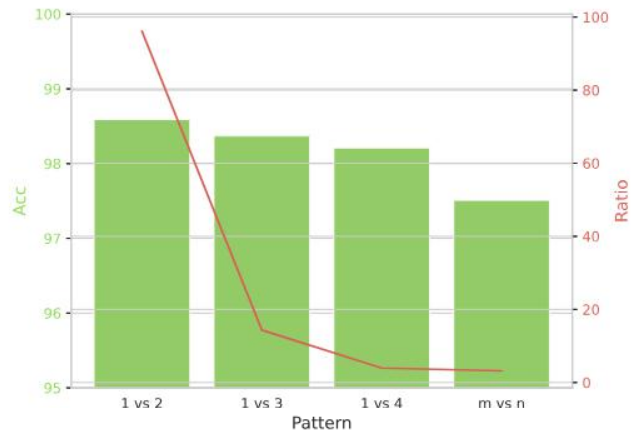


Figure 2: Accuracy and Appearance Ratio of Mapping Patterns in Simplified-to-Traditional Chinese Character Conversion.

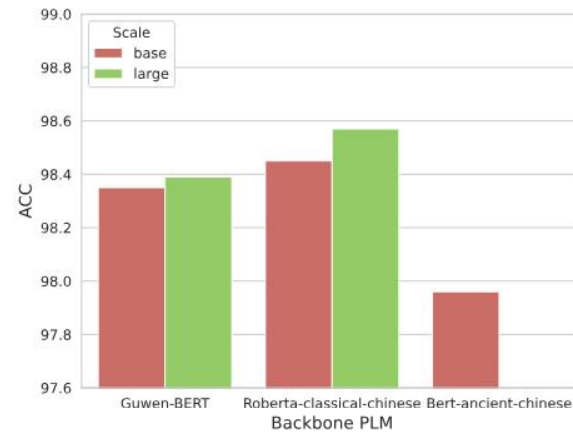


Figure 3: Accuracy of different backbone PLMs for simplified to traditional Chinese conversion. Both large and base models are included, except for “Bert-ancient-Chinese”.

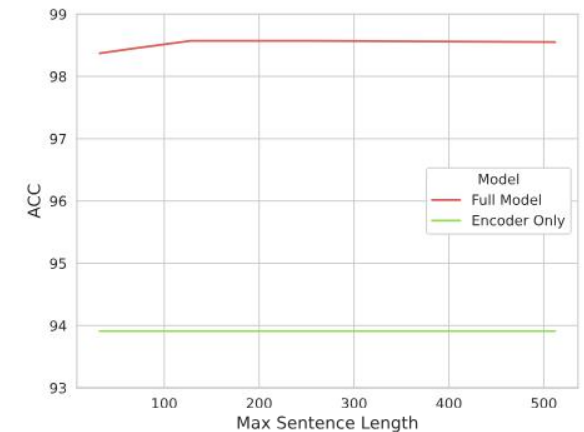


Figure 4: Accuracy of Proposed Model and Encoder-Only Baseline for Different Maximum Sentence Lengths in Simplified-to-Traditional Chinese Character Conversion.

## Case Study

Simplified	Traditional	Proposal	zhconv	aies.cn
<p>故国有贤君，折冲万里。</p> <p>中外若一，事无表里。</p> <p>文貌情用，相为内外表里。</p> <p>表里相资，古今一也。</p>	<p>故國有賢君，折沖萬里。</p> <p>中外若一，事無表裏。</p> <p>文貌情用，相為內外表裡。</p> <p>表里相資，古今一也。</p>	<b>里</b>	裡	裏
		<b>裏</b>	裡	裏
		<b>裏</b>	裡	裏
		<b>裏</b>	裡	裏
<p>星莫大于大辰，北斗常星。</p> <p>朝有变色之言，则下有争斗之患。</p>	<p>星莫大於大辰，北斗常星。</p> <p>朝有變色之言，則下有爭鬥之患。</p>	<b>斗</b>	鬥	鬥
		<b>鬥</b>	鬥	鬥
<p>百官饥饿，河内太守张杨使数千人负米贡饷。</p> <p>四时不出，天下大饥。</p>	<p>百官飢餓，河內太守張楊使數千人負米貢餉。</p> <p>四時不出，天下大饑。</p>	<b>飢</b>	飢	饑
		<b>饑</b>	飢	饑

Figure 5: Examples of Simplified-to-Traditional Chinese Character Conversion with **Incorrect Predictions** Highlighted in **Red**. The comparison targets are in bold. The translations are provided in the Appendix.



3

# Conclusion

---

## Conclusion

---

- We propose an unsupervised adaptive context-aware model for the simplified-traditional Chinese character convention task.
- To alleviate the one-to-many problem, we propose to introduce PLM for contextual semantic modeling in a reconstruction decoder.
- Based on the observation that different characters may require different levels of semantic modeling, we propose to apply early exit mechanism for inference.



That's all.  
Thank you.