# Correcting Chinese Word Usage Errors for Learning Chinese as a Second Language

Yow-Ting Shiue[1], Hen-Hsen Huang[1], and Hsin-Hsi Chen[1,2]

[1]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan

[2]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taipei, Taiwan

orina1123@gmail.com, hhhuang@nlg.csie.ntu.edu.tw, hhchen@ntu.edu.tw

## *Abstract*

With more and more people around the world learning Chinese as a second language, the need of Chinese error correction tools is increasing. In the HSK dynamic composition corpus, **word usage error (WUE)** is the **most common** error type. In this paper, we build a neural network model that **considers both target erroneous token and context** to generate a correction vector and compare it against a candidate vocabulary to propose suitable corrections. To deal with potential alternative corrections, the top five proposed candidates are judged by native Chinese speakers. For **more than 91%** of the cases, our system can propose **at least one acceptable correction within a list of five candidates**. To the best of our knowledge, this is the **first research** addressing **general-type Chinese WUE correction**. Our system can help non-native Chinese learners revise their sentences by themselves.

## Introduction

- **Chinese word usage error (WUE)**: incorrect token involving morphological, syntactical, or semantical problems
  - Incorrect word form e.g. *決解 ju´e ji˘e
  - Correct existent word that is improper for its context
- Goal: given a known erroneous token in a sentence segment → generate suitable correction
- Criteria for suitable correction
  1. **Correctness (C)**: result is syntactically and semantically correct
  2. **Similarity (S)**: meaning close to writer's intended meaning

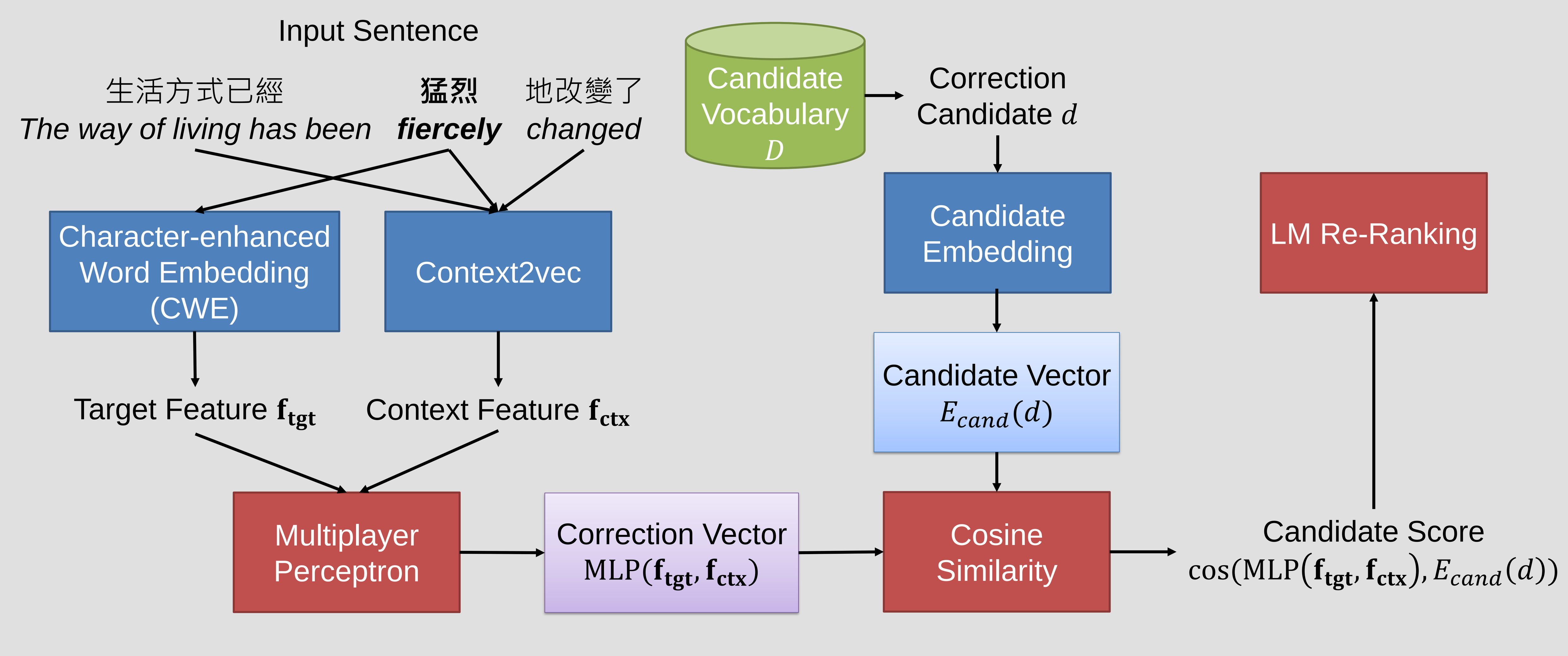
| | | C | S |
|---|---|---|---|
| | *生活方式已經**猛烈**地改變了 ( The way of living has been **fiercely** changed. ) | | |
| | *生活方式已經**暴烈**地改變了 ( ... has been **overpoweringly** changed. ) | X | O |
| | 生活方式已經**緩慢**地改變了 ( ... has been **slowly** changed. ) | O | X |
| V | 生活方式已經**劇烈**地改變了 ( ... has been **dramatically** changed. ) | O | O |

| | | C | S |
|---|---|---|---|
| | *發生這種情況的**情緒**很多 ( Many **emotions** happen this situation. ) | | |
| V | 發生這種情況的**因素**很多 ( Many **factors** can lead to this situation. ) | O | ? |

→ **C > S**: incorrect sentence can confuse language learner!

- Main contributions
  1. First study to correct all types of WUEs
  2. Human evaluation → at least one acceptable correction within top five candidates for more than 91% cases
  3. HSK WUE dataset with additional human annotations

## Neural Network-based Correction Generation Model




## Input Features

### Target CWE+P Word Embedding ($\text{CWE}_\text{w}$)

word; character: s, m, e

農產品 = 農產品 + 農 + 產 + 品

*農作品 = 農 + 作 + 品

解決 = 解決 + 解 + 決

### CWE Position-Insensitive Character Embedding ($\text{CWE}_\text{c}$)

*決解 = 決 + 決 + 決 + 解 + 解 + 解

### Context2vec Features

- 可是 每 個 人 的 [ 境況 ] 都 千差萬別 ($\text{C2V}_\text{tgt}$)
- $\text{C2V}_\text{ctx}$ = LSTM(可是 每 個 人 的) ⊕LSTM(千差萬別 都)

### POS Features

- 70% POS unchanged after correction
- Systematic changes e.g. NN ↔ VV
- One-hot encoding as input features

## LM Re-ranking

| | | C | S |
|---|---|---|---|
| | *到山頂**之間**路走得不容易 ( The road **between** the hilltop was not easy to walk. ) | | |
| | *到山頂**期間**路走得不容易 ( The road **period** the hilltop ... ) | X | O |
| V | 到山頂**的**路走得不容易 ( The road **to** the hilltop ... ) | O | O |

- Model sometimes generates candidates that seriously violate correctness criterion
- LM probability reflects the level of correctness

$$r_{\text{com}} = \frac{1}{\frac{\alpha}{r_{\text{LM}}} + \frac{1-\alpha}{r_{\text{DNN}}}}$$

- $r_{\text{com}}$ can be interpreted as rank, smaller better

## Evaluation

| Target features | Context features | Acc. | MRR | Hit@5 |
|---|---|---|---|---|
| **Baselines (No training on the WUE dataset)** | | | | |
| - | N-gram LM | 0.1659 | 0.2438 | 0.3268 |
| - | RNNLM | 0.1468 | 0.2208 | 0.2847 |
| - | $\text{C2V}_\text{ctx}$ | 0.0714 | 0.1170 | 0.1575 |
| **Correction Generation Model – C2V** | | | | |
| $\text{C2V}_\text{tgt}$ | - | 0.2507 | 0.3030 | 0.3561 |
| - | $\text{C2V}_\text{ctx}$ | 0.1249 | 0.1746 | 0.2273 |
| $\text{C2V}_\text{tgt}$ | $\text{C2V}_\text{ctx}$ | 0.3249 | 0.3891 | 0.4566 |
| **Correction Generation Model – CWE + Others** | | | | |
| $\text{CWE}_\text{w}$ | | 0.2898 | 0.3545 | 0.4195 |
| + $\text{CWE}_\text{c}$ | | 0.2946 | 0.3570 | 0.4234 |
| + $\text{C2V}_\text{tgt}$ | + $\text{C2V}_\text{ctx}$ | 0.3512 | 0.4250 | 0.5024 |
| + POS | | **0.3717** | **0.4378** | **0.5063** |

### Effect of LM Re-ranking

| Model | Acc. | MRR | Hit@5 | Hit@10 |
|---|---|---|---|---|
| Best MLP | 0.3717 | 0.4378 | 0.5063 | 0.5688 |
| + N-gram LM | **0.3727** | **0.4605** | **0.5561** | **0.6439** |
| + RNNLM | **0.3727** | 0.4527 | 0.5278 | 0.6205 |

### Human Evaluation

| Evaluation | Acc. | MRR | Hit@5 | Hit@10 |
|---|---|---|---|---|
| Ground-truth | 0.3727 | 0.4605 | 0.5561 | 0.6439 |
| + Annotation | 0.6829 | 0.7784 | 0.9122 | 0.9171 |

### Performance on most frequent POS tags

| POS (#) | Acc. | MRR | Hit@5 | Mean rank |
|---|---|---|---|---|
| VV (316) | 0.67 | 0.77 | 0.91 | 26.12 |
| NN (277) | 0.64 | 0.73 | 0.88 | 73.97 |
| AD (130) | 0.65 | 0.75 | 0.88 | 96.16 |
| P (62) | 0.81 | 0.88 | 0.95 | 3.10 |
| VA (45) | 0.60 | 0.76 | 0.98 | 1.98 |
| DEV (23) | 1.00 | 1.00 | 1.00 | 1.00 |
| PN (21) | 0.71 | 0.80 | 0.95 | 2.33 |

## Conclusion

- MLP correction generation model considers both target & context
- Apply LM re-ranking to emphasize correctness
- Future work: phonetical similarity e.g. (影響 y˘ıng xi˘ang,印象 y`ın xi`ang)