# Detection of Chinese Word Usage Errors for Non-Native Chinese Learners with Bidirectional LSTM

Yow-Ting Shiue, Hen-Hsen Huang and Hsin-Hsi Chen
Department of Computer Science and Information Engineering
National Taiwan University, Taipei, Taiwan
orina1123@gmail.com, hhhuang@nlg.csie.ntu.edu.tw, hhchen@ntu.edu.tw

## Abstract

Selecting appropriate words to compose a sentence is one common problem faced by non-native Chinese learners. In this paper, we propose (bidirectional) LSTM sequence labeling models and explore various features to detect word usage errors in Chinese sentences. By combining CWINDOW word embedding features and POS information, the best bidirectional LSTM model achieves accuracy 0.5138 and MRR 0.6789 on the HSK dataset. For 80.79% of the test data, the model ranks the ground-truth within the top two at position level.

## Introduction

**Chinese word usage error (WUE)**
Grammatically or semantically incorrect token
- Written in a wrong form
- Existent but is improper for its context

Many Chinese WUEs result from subtle **semantic** unsuitability instead of violation of syntactic constraints

(E1) 人們有(*權力,權利)吃安全的食品。
(People have the (*power, right) to enjoy safe food.)

- Both 權力(power) and 權利(right) are existent nouns in Chinese
- Both versions are grammatically correct

→ Difficult to formulate an explicit rule for this kind of errors

## WUE Detection Based on Bidirectional LSTM

Chinese WUE detection → **sequence labeling problem**
- Each token is labeled either correct (0) or incorrect (1)
- **LSTM**: capture long dependencies among time steps
  → suitable for modeling complex dependencies of the erroneous token on other parts of the sentence
- **Bidirectional LSTM**: forward + backward LSTM layer

(E3) 店是爸爸(*留在,留給)我們的。
(The store is our father left (*at,to) us.)

- Need future information to detect the error

## Sequence Embedding

**Word Embeddings** [trainable] [dim = 400]
- Random
- CBOW/Skip-gram
- CWINDOW/Structured Skip-gram
  - Consider context word **order**
  - **CWIN**: concatenate context word vectors
  - **Struct-SG**: different projection matrices for context words in different relative position with target word

**POS Embeddings** [trainable] [dim = 20] (# unique POS = 30)
- Random

## Token Features

- Derived from Google Chinese Web 5-gram corpus [external info.]

**Out-of-Vocabulary Indicator**
0 / 1

**N-gram Probability Features**
- Compute 2gram & 3gram probability of each token using occurrence count
- How likely an expression is valid

## Experiments

**Dataset**
- Each sentence segment has **exactly one** erroneous token
- 10,510 sentence segments
  - Train/val/test = 8:1:1

**Evaluation**
- **Accuracy**
  - Strict, avg. segment len.: 9.24
- **MRR**
- **Hit@2 Rate**

Collocation error, involve a **pair** of words
(E2) * 學習的知識也很差
(The knowledge learned is also very bad.)
Both corrections acceptable:
(E4) 學習的知識也很**不足**
(The knowledge learned is also **insufficient**.)
(E5) 學習的**態度**也很差
(The **attitude** of learning is also very bad.)
- **Hit@20% Rate**

Regard one test instance as correct if the answer is ranked within the top $\max(1, \lfloor len * 20\% \rfloor)$ candidates
  - Different level of difficulty according to $len$ (# tokens)

| Model | Features | Accuracy | MRR | Hit@2 | Hit@20% |
|---|---|---|---|---|---|
| Rand. baseline | - | 0.1239 | 0.3312 | 0.2478 | 0.1611 |
| LSTM | Rand. Emb. | 0.4186 | 0.6010 | 0.7222 | 0.6565 |
| | CBOW | 0.4072 | 0.5923 | 0.7155 | 0.6432 |
| | SG | 0.4072 | 0.5910 | 0.7146 | 0.6365 |
| | CWIN | 0.4853 | 0.6537 | 0.7774 | 0.7031 |
| | Struct-SG | 0.4710 | 0.6412 | 0.7650 | 0.6889 |
| Bi-LSTM | CWIN | 0.4795 | 0.6547 | 0.7840 | 0.7174 |
| | + POS | **0.5138** | **0.6789** | 0.8097 | 0.7479 |
| | + N-gram | 0.4948 | 0.6719 | **0.8173** | **0.7507** |

**Table 1** LSTM/Bi-LSTM with different sets of features

| Length | (# tests) | # proposed | LSTM | Bi-LSTM |
|---|---|---|---|---|
| **< 10** | (645) | 1 | 0.7426 | **0.7659** |
| **10 ~ 14** | (317) | 2 | 0.6908 | **0.7319** |
| **15+** | (89) | 3+ | **0.7416** | 0.7079 |

**Table 2** Hit@20% rates on segments with different lengths

| | |
|---|---|
| **# correct ($c_1 = a$)** | 520 (49.48%) |
| **# tests where $c_2 = a$** | 339 (32.25%) |
| **Avg. $dis(c_1, c_2)$ when $c_2 = a$** | 2.07 |
| **# tests where $c_2 = a$ and $dis(c_1, c_2) = 1$** | 129 (12.27%) |

**Table 3** Dependency distance analysis
- $a$: ground-truth error position
- $c_1$, $c_2$: first and second candidate positions proposed by model
- $dis(c_1, c_2)$: shortest path distance between $c_1$ and $c_2$ on undirected dependency graph

## Conclusion

- LSTM-based sequence labeling model for detecting WUEs in sentences written by non-native Chinese learners
- CWIN/Struct-SG are better word features
- Bi-LSTM > LSTM
- Best model can rank ground-truth error position within **top two** in **80.97%** of the cases