Overview of the NLP-TEA 2015 Shared Task for Chinese Grammatical Error Diagnosis

Lung-Hao Lee, National Taiwan Normal University
Liang-Chih Yu, Yuan Ze University
Li-Ping Chang, National Taiwan Normal University
Introduction

• The NLP-TEA 2015 shared task features a Chinese Grammatical Error Diagnosis (CGED) task, providing an evaluation platform for the development and implementation of NLP tools for computer-assisted Chinese learning.
Shared Task Description

• The developed tool is expected to identify the error types and its position at which it occurs in the sentence.

• Four PADS error types are included in the target modification taxonomy:
  – Mis-Ordering (Permutation)
  – Redundancy (Addition)
  – Omission (Deletion)
  – Mis-Selection (Substitution)

• For the sake of simplicity, the input sentence is selected to contain one defined error types.
Testing Examples

• Example 1
  Input: (sid=B2-0080) 他是我的以前的室友
  Output: B2-0080, 4, 4, Redundant

• Example 2
  Input: (sid=A2-0017)  那電影是機器人的故事
  Output: A2-0017, 2, 2, Missing

• Example 3
  Input: (sid=A2-0017)  那部電影是機器人的故事
  Output: A2-0017, Correct

• Example 4
  Input: (sid=B1-1193)  吳先生是修理腳踏車的拿手
  Output: B1-1193, 11, 12, Selection

• Example 5
  Input: (sid=B2-2292)  所以我不會讓失望她
  Output: B2-2292, 7, 9, Disorder
Data Preparation

• The essay section of the computer-based Test of Chinese as a Foreign Language (TOCFL)

• Native Chinese speakers were trained to manually annotate grammatical errors and provide corrections corresponding to each error
Training Set

• This set included 2,205 selected sentences
• Error types were categorized as redundant (430 instances), missing (620), selection (849), and disorder (306)
• Each sentence is represented in SGML format

```
<DOC>
  <SENTENCE id="B1-1120">
    我的中文進步了非常快
  </SENTENCE>
  <MISTAKE start_off="7" end_off="7">
    <TYPE>
      Selection
    </TYPE>
    <CORRECTION>
      我的中文進步得非常快
    </CORRECTION>
  </MISTAKE>
</DOC>
```
Dryrun Set

- A total of 55 sentences were given to participants to familiarize themselves with the final testing process.
- The purpose is output format validation only.
- No matter which performance can be achieved that will not be included in our official evaluation.

NLP-TEA 2015 @ Beijing, China
Test Set

• This set consists of 1,000 testing sentences

• Half of these sentences contained no grammatical errors, while the other half included a single defined grammatical error: redundant (132 instances), missing (126), selection (110), and disorder (132)
Performance Metrics

• Correctness is determined at three levels
  – Detection-level
  – Identification-level
  – Position-level

• Metrics
  – False positive rate (FPR) = \( \frac{FP}{FP+TP} \)
  – Accuracy = \( \frac{TP+TN}{TP+FP+TN+FN} \)
  – Precision = \( \frac{TP}{TP+FP} \)
  – Recall = \( \frac{TP}{TP+FN} \)
  – \( F1 \) = \( \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \)
Evaluation Examples

• FPR = 0.5
• Detection-level Acc. = 0.75, Pre.=0.67, Rec.=1, F1=0.8
• Correction-level Acc. = 0.625, Pre.=0.5, Rec.=0.75, F1=0.6
• Position-level Acc. = 0.5, Pre.=0.33, Rec.=0.5, F1=0.4

• System Results

• Gold Standard
13 Participants and 18 Submitted Runs

<table>
<thead>
<tr>
<th>Participant (Ordered by abbreviations of names)</th>
<th>#Runs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adam Mickiewicz University on Poznan (AMU)</td>
<td>0</td>
</tr>
<tr>
<td>University of Cambridge (CAM)</td>
<td>0</td>
</tr>
<tr>
<td>Chinese Academy of Sciences (CAS)</td>
<td>0</td>
</tr>
<tr>
<td>Confucius Institute of Rutgers University (CIRU)</td>
<td>0</td>
</tr>
<tr>
<td>Chaoyang University of Technology (CYUT)</td>
<td>3</td>
</tr>
<tr>
<td>Harbin Institute of Technology Shenzhen Graduate School (HITSZ)</td>
<td>3</td>
</tr>
<tr>
<td>Lingage Inc. (Lingage)</td>
<td>0</td>
</tr>
<tr>
<td>National Chiayi University (NCYU)</td>
<td>3</td>
</tr>
<tr>
<td>National Taiwan Ocean University (NTOU)</td>
<td>3</td>
</tr>
<tr>
<td>National Taiwan University (NTU)</td>
<td>0</td>
</tr>
<tr>
<td>South China Agriculture University (SCAU)</td>
<td>3</td>
</tr>
<tr>
<td>Tokyo Metropolitan University (TMU)</td>
<td>3</td>
</tr>
<tr>
<td>University of Leeds (UL)</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>18</strong></td>
</tr>
</tbody>
</table>
### Testing Results

<table>
<thead>
<tr>
<th>Submission</th>
<th>False Positive Rate</th>
<th>Detection Level</th>
<th>Identification Level</th>
<th>Position Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYUT-Run1</td>
<td>0.096</td>
<td>0.584</td>
<td>0.7333</td>
<td>0.264</td>
</tr>
<tr>
<td>CYUT-Run2</td>
<td><strong>0.082</strong></td>
<td>0.579</td>
<td><strong>0.7453</strong></td>
<td>0.24</td>
</tr>
<tr>
<td>CYUT-Run3</td>
<td>0.132</td>
<td>0.579</td>
<td>0.6872</td>
<td>0.29</td>
</tr>
<tr>
<td>HITSZ-Run1</td>
<td>0.956</td>
<td>0.509</td>
<td>0.5047</td>
<td>0.974</td>
</tr>
<tr>
<td>HITSZ-Run2</td>
<td>0.938</td>
<td>0.505</td>
<td>0.5027</td>
<td>0.948</td>
</tr>
<tr>
<td>HITSZ-Run3</td>
<td>0.884</td>
<td>0.51</td>
<td>0.5056</td>
<td>0.904</td>
</tr>
<tr>
<td>NCYU-Run1</td>
<td>0.48</td>
<td>0.53</td>
<td>0.5294</td>
<td>0.54</td>
</tr>
<tr>
<td>NCYU-Run2</td>
<td>0.396</td>
<td>0.567</td>
<td>0.5724</td>
<td>0.53</td>
</tr>
<tr>
<td>NCYU-Run3</td>
<td>0.374</td>
<td>0.607</td>
<td>0.6112</td>
<td>0.588</td>
</tr>
<tr>
<td>NTOU-Run1</td>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>NTOU-Run2</td>
<td>0.914</td>
<td>0.531</td>
<td>0.5164</td>
<td>0.976</td>
</tr>
<tr>
<td>NTOU-Run3</td>
<td>0.948</td>
<td>0.519</td>
<td>0.5098</td>
<td>0.986</td>
</tr>
<tr>
<td>SCAU-Run1</td>
<td>0.62</td>
<td>0.505</td>
<td>0.504</td>
<td>0.63</td>
</tr>
<tr>
<td>SCAU-Run2</td>
<td>0.636</td>
<td>0.503</td>
<td>0.5023</td>
<td>0.642</td>
</tr>
<tr>
<td>SCAU-Run3</td>
<td>0.266</td>
<td>0.503</td>
<td>0.5056</td>
<td>0.272</td>
</tr>
<tr>
<td>TMU-Run1</td>
<td>0.478</td>
<td>0.516</td>
<td>0.5162</td>
<td>0.51</td>
</tr>
<tr>
<td>TMU-Run2</td>
<td>0.134</td>
<td>0.524</td>
<td>0.5759</td>
<td>0.182</td>
</tr>
<tr>
<td>TMU-Run3</td>
<td>0.35</td>
<td>0.546</td>
<td>0.5581</td>
<td>0.442</td>
</tr>
</tbody>
</table>
Summary

• It is a really difficult task to develop the computer-assisted Chinese learning tool, since there are only target sentences without the help of their context.

• None of the submitted systems provided superior performance. In general, this research problem still has long way to go.
Conclusions

• All submissions contribute to the common effort to produce an effective Chinese grammatical diagnosis tool

• The individual reports in the shared task proceedings provide useful insight into Chinese language processing
Future Work

• **NLP-TEA-3 Workshop in COLING 2016**
  – To be bided
  – Osaka, Japan

• **The Shared Task**
  – Chinese Grammatical Error Diagnosis
Acknowledgments

• Ministry of Education, Taiwan
  – Aim for the Top University Project
  – Center of Learning Technology for Chinese, NTNU
  – Innovative Center for Big Data and Digital Convergence, YZU

• Ministry of Science and Technology, Taiwan
  – International Research-Intensive Center of Excellence Program
  – Grant no.: MOST 104-2911-I-003-301
THANK YOU

• All data sets with gold standards and evaluation tool are publicly available for research purposes at

http://ir.itc.ntnu.edu.tw/lre/nlptea15cged.htm