Overview of SIGHAN 2015 Bake-off for Chinese Spelling Check

Yuen-Hsien Tseng (曾元顯), National Taiwan Normal Univ.
Lung-Hao Lee (李龍豪), National Taiwan Normal Univ.
Li-Ping Chang (張莉萍), National Taiwan Normal Univ.
Hsin-Hsi Chen (陳信希), National Taiwan Univ.

Introduction

- Chinese spelling checkers are difficult
  - No word delimiters exist among Chinese words
  - A Chinese word can contain only a single character or multiple characters
  - More than 13 thousand characters

- The spelling checker is expected to identify all possible spelling errors, highlight their locations and suggest possible corrections

Chinese Spelling Check Evaluations

- The 1st Chinese Spelling Check Bake-off
  - Native Chinese speakers
  - SIGHAN-2013 workshop @ Nagoya, Japan
- The 2nd Chinese Spelling Check Bake-off
  - Chinese as a foreign language learners
  - CIPS-SIGHAN joint CLP-2014 conference @ Wuhan
- The 3rd Chinese Spelling Check Bake-off
  - Chinese as a foreign language learners
  - SIGHAN-2015 workshop @ Beijing, China

Task Description

- The input instance is given a unique passage number PID
- Each character or punctuation mark occupies 1 spot for counting location

- If the passage contains no spelling errors, the checker should return "PID, 0"
- If an input passage contains at least one spelling error, the output format is "PID, [, location, correction]+"

Testing Examples

- Example 1
  - Input: (pid=A2-0047-1) 我真的洗碗我可以去看你
  - Output: A2-0047-1, 4, 希, 5, 望
- Example 2
  - Input: (pid=B2-1670-2) 在日本，大學生打工的情況是相當普偏的。
  - Output: B2-1670-2, 17, 遍
- Example 3
  - Input: (pid=B2-1903-7) 我也是你的朋友，我會永遠在你身邊。
  - Output: B2-1903-7, 0 CORRECT

Data Preparation

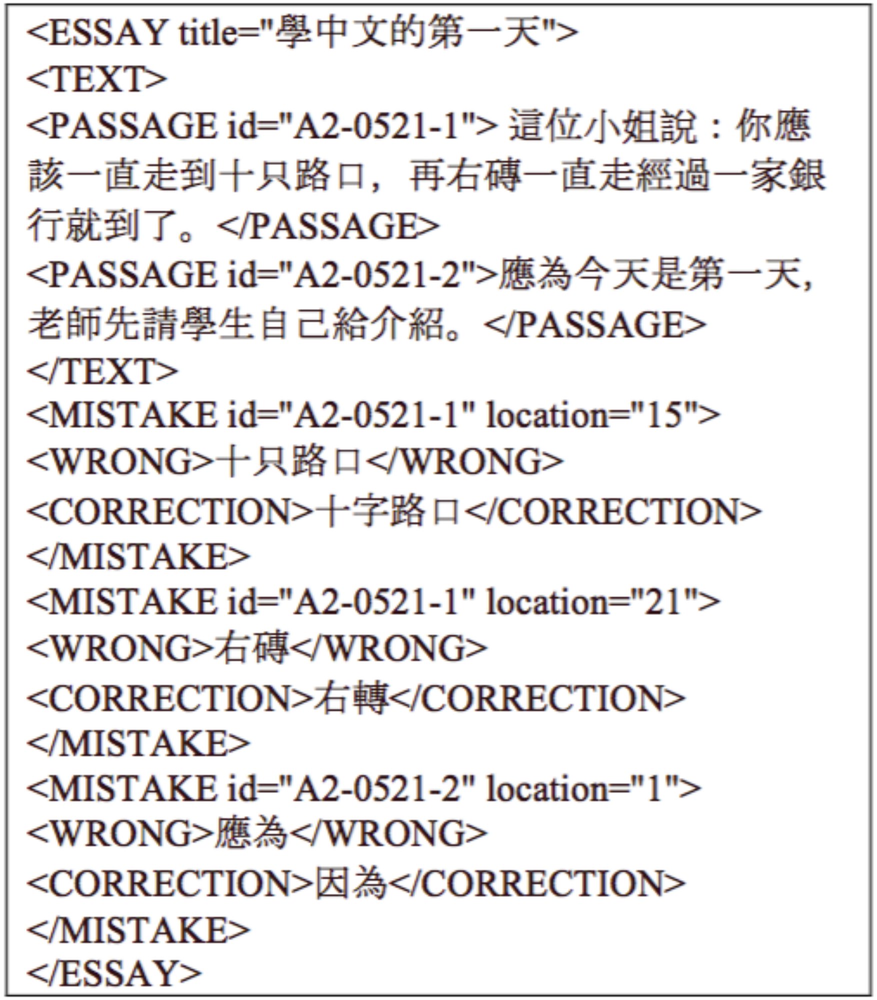
- The essay section of the computer-based Test of Chinese as a Foreign Language (TOCFL)

- The spelling errors were manually annotated by trained native Chinese speakers, who also provided corrections corresponding to each error.

Training Set

- This set included 970 selected essays with a total of 3,143 spelling errors.

- Each essay is shown in terms of SGML format



```
<ESSAY title="學中文的第一天">
<TEXT>
<PASSAGE id="A2-0521-1"> 這位小姐說：你應
該一直走到十只路口，再右磚一直走經過一家銀
行就到了。</PASSAGE>
<PASSAGE id="A2-0521-2">應為今天是第一天，
老師先請學生自己給介紹。</PASSAGE>
</TEXT>
<MISTAKE id="A2-0521-1" location="15">
<WRONG>十只路口</WRONG>
<CORRECTION>十字路口</CORRECTION>
</MISTAKE>
<MISTAKE id="A2-0521-1" location="21">
<WRONG>右磚</WRONG>
<CORRECTION>右轉</CORRECTION>
</MISTAKE>
<MISTAKE id="A2-0521-2" location="1">
<WRONG>應為</WRONG>
<CORRECTION>因為</CORRECTION>
</MISTAKE>
</ESSAY>
```

Dryrun Set

- A total of 39 passages were given to participants to familiarize themselves with the final testing process.

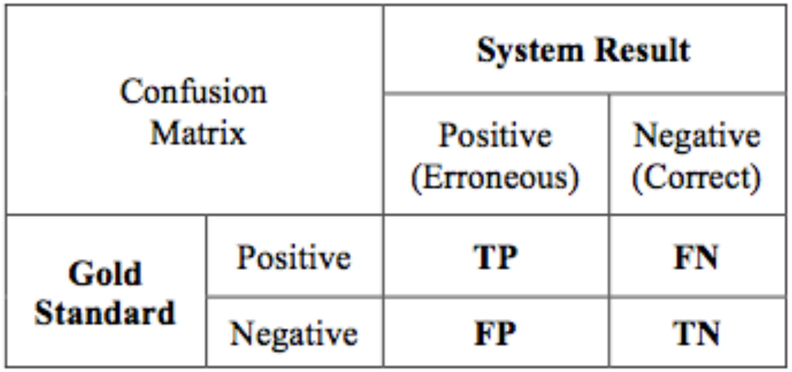
- The purpose is to validate the submitted output format only, and no dryrun outcomes were considered in the official evaluation

Test Set

- This set consists of 1,100 testing passages. Half of these passages contained no spelling errors, while the other half included at least one spelling error
- Open test policy: employing any linguistic and computational resources to detect and correct spelling errors are allowed.

Performance Metrics

- Correctness is determined at two levels
  - Detection-level
  - Correction-level



Confusion Matrix		**System Result**	
		Positive (Erroneous)	Negative (Correct)
Gold Standard	Positive	**TP**	**FN**
	Negative	**FP**	**TN**

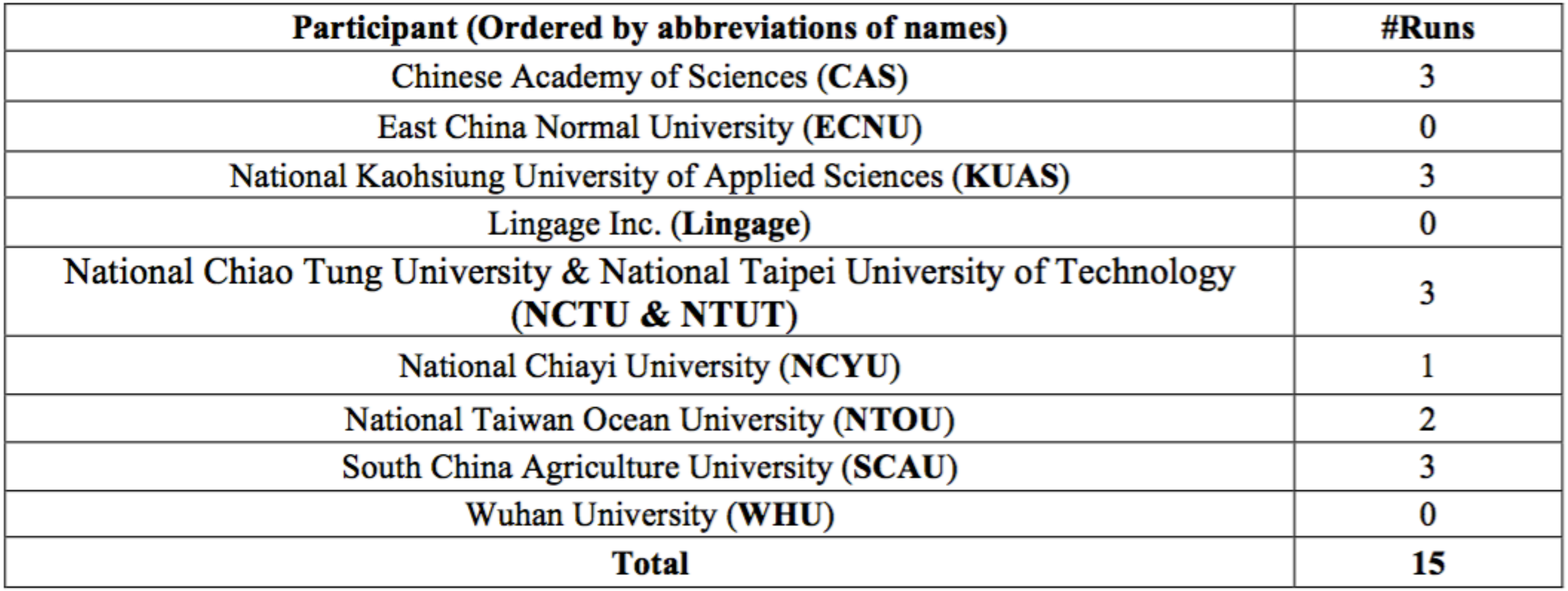
- Metrics
  - False positive rate (FPR) = FP / (FP+TP)
  - Accuracy = (TP+TN) / (TP+FP+TN+FN)
  - Precision = TP / (TP+FP)
  - Recall = TP / (TP+FN)
  - F1 = 2 * Precision * Recall / (Precision+Recall)

Evaluation Examples

- System Results: “A2-0092-2, 5, 玩”, “A2-0243- 1, 3, 件, 4, 康”, “B2-1923-2, 8, 誤, 41, 情”, “B2- 2731-1, 0”, and “B2-3754-3, 11, 觀”
- Gold Standard: “A2-0092-2, 0”, “A2-0243-1, 3, 健, 4, 康”, “B2-1923-2, 8, 誤, 41, 情”, “B2-2731-1, 0”, and “B2-3754-3, 10, 觀”,

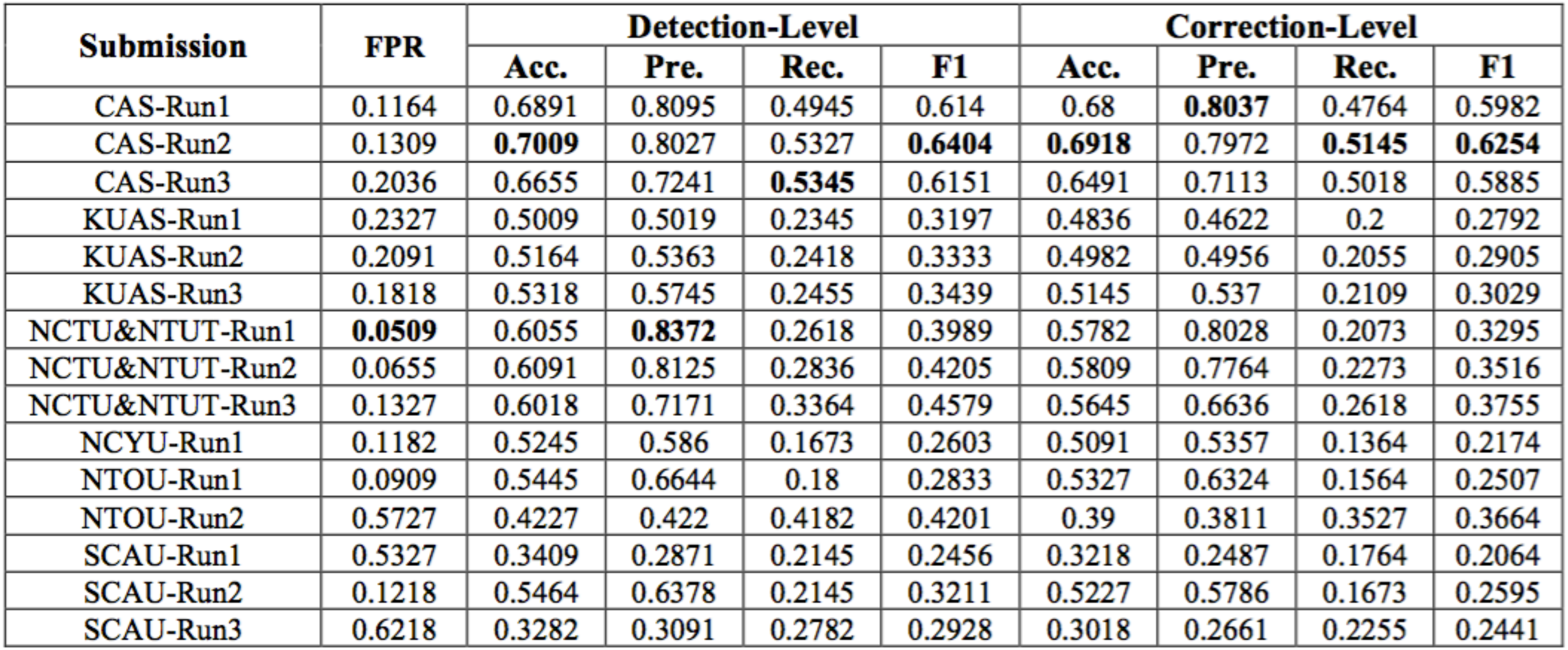
- FPR = 0.5
- Detection-level Acc. = 0.6, Pre.=0.5, Rec.=0.67, F1=0.57
- Correction-level Acc. = 0.4, Pre.=0.25, Rec.=0.33, F1=0.28

9 Participants & 15 Runs



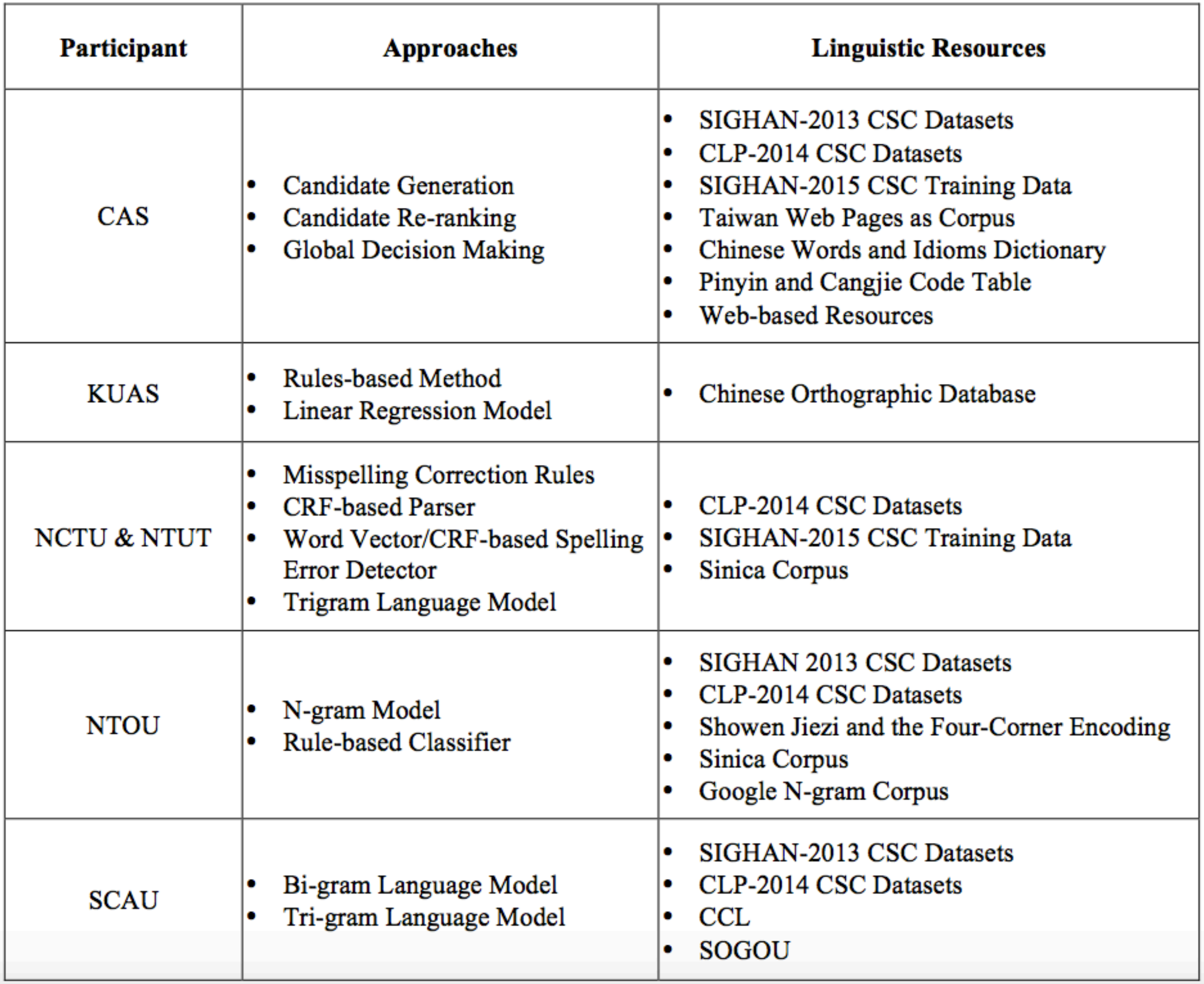
Participant (Ordered by abbreviations of names)	**#Runs**
Chinese Academy of Sciences (**CAS**)	3
East China Normal University (**ECNU**)	0
National Kaohsiung University of Applied Sciences (**KUAS**)	3
Lingage Inc. (**Lingage**)	0
National Chiao Tung University & National Taipei University of Technology (**NCTU & NTUT**)	3
National Chiayi University (**NCYU**)	1
National Taiwan Ocean University (**NTOU**)	2
South China Agriculture University (**SCAU**)	3
Wuhan University (**WHU**)	0
Total	**15**

Testing Results



Submission	FPR	Detection-Level				Correction-Level			
		Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1
CAS-Run1	0.1164	0.6891	0.8095	0.4945	0.614	0.68	**0.8037**	0.4764	0.5982
CAS-Run2	0.1309	**0.7009**	0.8027	0.5327	**0.6404**	**0.6918**	0.7972	**0.5145**	**0.6254**
CAS-Run3	0.2036	0.6655	0.7241	**0.5345**	0.6151	0.6491	0.7113	0.5018	0.5885
KUAS-Run1	0.2327	0.5009	0.5019	0.2345	0.3197	0.4836	0.4622	0.2	0.2792
KUAS-Run2	0.2091	0.5164	0.5363	0.2418	0.3333	0.4982	0.4956	0.2055	0.2905
KUAS-Run3	0.1818	0.5318	0.5745	0.2455	0.3439	0.5145	0.537	0.2109	0.3029
NCTU&NTUT-Run1	**0.0509**	0.6055	**0.8372**	0.2618	0.3989	0.5782	0.8028	0.2073	0.3295
NCTU&NTUT-Run2	0.0655	0.6091	0.8125	0.2836	0.4205	0.5809	0.7764	0.2273	0.3516
NCTU&NTUT-Run3	0.1327	0.6018	0.7171	0.3364	0.4579	0.5645	0.6636	0.2618	0.3755
NCYU-Run1	0.1182	0.5245	0.586	0.1673	0.2603	0.5091	0.5357	0.1364	0.2174
NTOU-Run1	0.0909	0.5445	0.6644	0.18	0.2833	0.5327	0.6324	0.1564	0.2507
NTOU-Run2	0.5727	0.4227	0.422	0.4182	0.4201	0.39	0.3811	0.3527	0.3664
SCAU-Run1	0.5327	0.3409	0.2871	0.2145	0.2456	0.3218	0.2487	0.1764	0.2064
SCAU-Run2	0.1218	0.5464	0.6378	0.2145	0.3211	0.5227	0.5786	0.1673	0.2595
SCAU-Run3	0.6218	0.3282	0.3091	0.2782	0.2928	0.3018	0.2661	0.2255	0.2441

A Summary of Developed Systems



Participant	Approaches	Linguistic Resources
CAS	• Candidate Generation • Candidate Re-ranking • Global Decision Making	• SIGHAN-2013 CSC Datasets • CLP-2014 CSC Datasets • SIGHAN-2015 CSC Training Data • Taiwan Web Pages as Corpus • Chinese Words and Idioms Dictionary • Pinyin and Cangjie Code Table • Web-based Resources
KUAS	• Rules-based Method • Linear Regression Model	• Chinese Orthographic Database
NCTU & NTUT	• Misspelling Correction Rules • CRF-based Parser • Word Vector/CRF-based Spelling Error Detector • Trigram Language Model	• CLP-2014 CSC Datasets • SIGHAN-2015 CSC Training Data • Sinica Corpus
NTOU	• N-gram Model • Rule-based Classifier	• SIGHAN 2013 CSC Datasets • CLP-2014 CSC Datasets • Showen Jiezi and the Four-Corner Encoding • Sinica Corpus • Google N-gram Corpus
SCAU	• Bi-gram Language Model • Tri-gram Language Model	• SIGHAN-2013 CSC Datasets • CLP-2014 CSC Datasets • CCL • SOGOU

Conclusions and Future Work

- All submissions contribute to the knowledge in search for an effective Chinese spell checkers

- The individual reports in the Bake-off proceedings provide useful insight into Chinese language processing

- The future direction focuses on the development of Chinese grammatical error correction

Acknowledgments

- National Taiwan Normal University
- Ministry of Education, Taiwan
  - Aim for the Top University Project
  - Center of Learning Technology for Chinese
- 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/sighan8csc.html