Chinese Grammatical Error Diagnosis System Based on Hybrid Model

Xiupeng Wu, Peijie Huang *, Jundong Wang, Qingwen Guo, Yuhong Xu, Chuping Chen

College of Mathematics and Informatics
South China Agricultural University
Guangzhou 510642, Guangdong, China
zxc2012@gmail.com, pjhuang@scau.edu.cn, mo_xiao_wang@163.com,
tryven.guo@qq.com, 137610184@qq.com, 568093091@qq.com

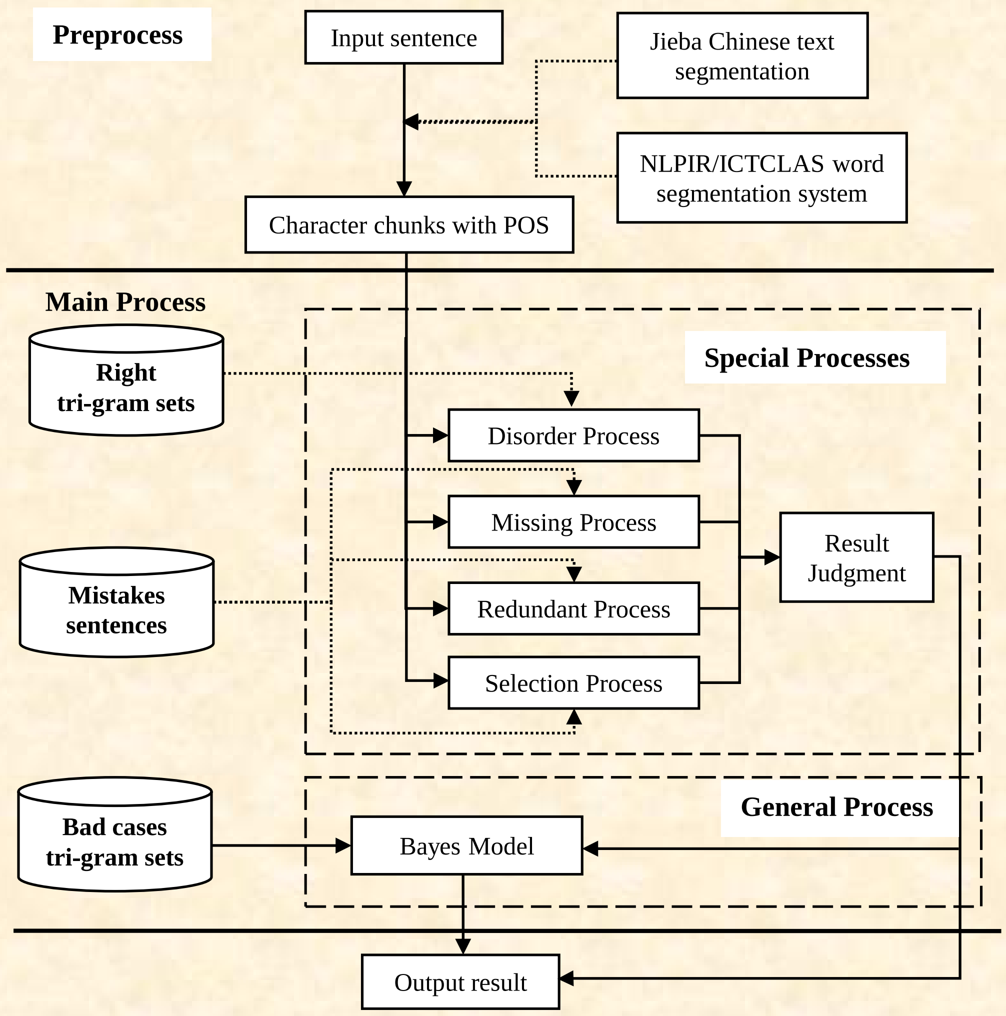

Figure 1. The flowchart of the hybrid CGED system.

	FPR	DA	DP	DR	DF	IA	IP	IR	IF	PA	PP	PR	PF
Run1	0.620	0.505	0.504	0.630	0.560	0.287	0.238	0.194	0.214	0.217	0.080	0.054	0.065
Run2	0.636	0.503	0.502	0.642	0.564	0.279	0.234	0.194	0.212	0.209	0.078	0.054	0.064
Run3	0.266	0.503	0.506	0.272	0.354	0.416	0.269	0.098	0.144	0.385	0.119	0.036	0.055
Average	0.538	0.534	0.560	0.607	0.533	0.335	0.329	0.208	0.233	0.263	0.166	0.064	0.085
Best	0.082	0.607	0.745	1.000	0.675	0.525	0.617	0.364	0.358	0.505	0.529	0.160	0.174

Figure 2. Evaluation results of NLP-TEA-2 CFL final test.

Disorder:
O:所以我不會讓失望她
Segmentation: 所以/c我/r 不會/v讓/v失望/v她/r
All Permutations:
所以我不會讓失望她
所以我不會讓她失望
(Highest probability in tri-gram)
所以我不會失望她讓
我不會讓失望她所以
所以我讓失望她不會
所以我讓失望不會她
所以不會讓失望她我
…
M:所以我不會讓她失望

Missing:
O: 我高興我的老師是那位小姐
Segmentation: 我/r高興/a我/r的/uj老師/n 是/v那位/r小姐/nr
Detection rule: “r+b+r+uj+n+v+r+nr”
Correction rule: “r+很+b+r+uj+n+v+r+nr”
M: 我很高興我的老師是那位小姐

Redundant:
O: 晚上五點半他們到了火車站去
Segmentation: 晚上/t五點半/m他們/r 到/v了/ul火車/n站/v去/v
Rule: (“v+ul+n”, “ v+n”, 2)
M: 晚上五點半他們到火車站去

Selection:
O: 你決定那個電影
Segmentation:你/rr決定/v那個/rz電影/n
Detection rule: .[^部]/mq+電影
Correction rule: .部/mq+電影
M: 你決定那部電影

Figure 3. Four examples of dealing with the different error.

Introduction

Chinese as a foreign language (CFL) is booming in recent decades. The number of (CFL) learners is expected to become larger for the years to come (Xiong et al., 2014). But the flexibility and complication in Chinese morphology, pronunciations and grammar make Chinese become one of the hardest languages to learn. If you cannot make good use of the grammatical rules, maybe the many different meaning or error meaning of the sentence will be get. This system proposes a hybrid model for CGED shared task by integrating rule-based methods and n-gram statistical methods to detect Chinese grammatical errors, identify the error type and point out the position of error in the input sentences.

System Overview

Figure 1 shows the flowchart of our CGED system. The system is mainly composed by two processes: preprocess and main process. In addition, main process contains two subprocesses: special process and general process. It performs CGED in the following steps:

1. Given a test sentence, the CGED system gets the character chunks in the sentence with POS. We uses “Jieba” Chinese text segmentation and NLPIR/ ICTCLAS Chinese text segmentation to achieve the goal.
2. For each chunk in this sentence, the system will enumerate every rule in the missing, redundant and selection rules sets. In the meanwhile, we got the all permutations of the chunks. What’s more we use tri-gram to calculate model the probabilities of each generated sentence in the all permutations and pop the highest one. We will get a candidate sentence set after this step.
3. If the candidate sentence set has only one sentence, the system will return related data based on the sentence. If not, system will carry out the general process.

Special processes and General process

Special processes contains four subprocesses: disorder process , missing process ,redundant process and selection process.

1. Disorder process: system calculates the probability of each sentence in the sentence set generated by tri-gram. If the highest probability one differs from the origin one, system judges that the sentence has disorder error.
2. Missing process , selection process and redundant process: through the collection of the grammar deletion, extract the sentence features of the deletion of the grammar, and analyze the grammar and summarize the relevant rules.

General process: By comparing the four non-standard corpora to the standard corpus, System extracts the four kinds of a list of ungrammatical n-grams list corresponding to the four kinds of non-standard corpora and treat them as key expressions. With these n-gram lists, we trained a classifiers n-gram based Multinomial Na ïve Bayes (MNB) to identify the grammatical error type.

Evaluation Results

The CGED task of CFL attracted 13 research teams. Among 13 registered teams, 6 participants submitted their testing results. For formal testing. Finally, there are 18 runs submitted in total. We use the 30% of NLP-TEA-1 CFL Datasets as validation set to test the effect and performance of the four special processes and the MNB. We found the approach using MNB has mostly no advantage with that only using special process. So, the three runs of our system submitted to NLP-TEA-2 CLF final test are all based on the four special processes. Figure 2 shows the evaluation results of the NLP-TEA-2 CFL final test. As we can see from Figure 2, we achieve a result close to the average level.

Selected References:

- Xiong J. H., Zhao Q., Hou J.P., et al. (2014). Extended HMM and Ranking Models for Chinese Spelling Correction. In Proceedings of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP 2014)
- Cheng, S. M., Yu, C. H., & Chen, H. H. (2014). Chinese Word Ordering Errors Detection and Correction for Non-Native Chinese Learners. In Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014) , Dublin, Ireland, pp. 279-289.
- Chodorow, M., Dickinson, M., Israel, R., & Tetreault, J. R. (2012). Problems in Evaluating Grammatical Error Detection Systems. In Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012), Mumbai, India, pp. 611-628