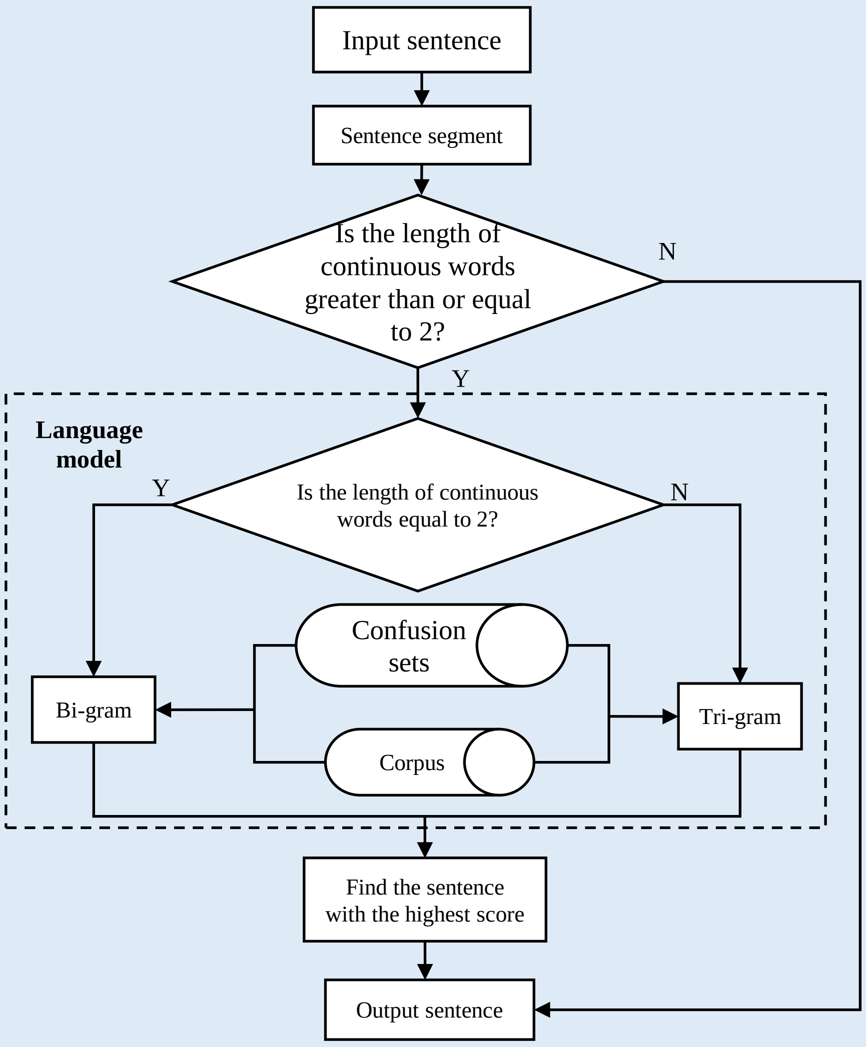
Chinese Spelling Check System Based on N-gram Model

Weijian Xie, Peijie Huang*, Xinrui Zhang, Kaiduo Hong, Qiang Huang,
Bingzhou Chen, Lei Huang
College of Mathematics and Informatics
South China Agricultural University
Guangzhou 510642, Guangdong, China
*pjhuang@scau.edu.cn



Input sentence

Sentence segment

Is the length of continuous words greater than or equal to 2?

N

Y

Language model

Y

Is the length of continuous words equal to 2?

N

Confusion sets

Bi-gram

Tri-gram

Corpus

Find the sentence with the highest score

Output sentence

Figure 1. The flowchart of the CSC system.

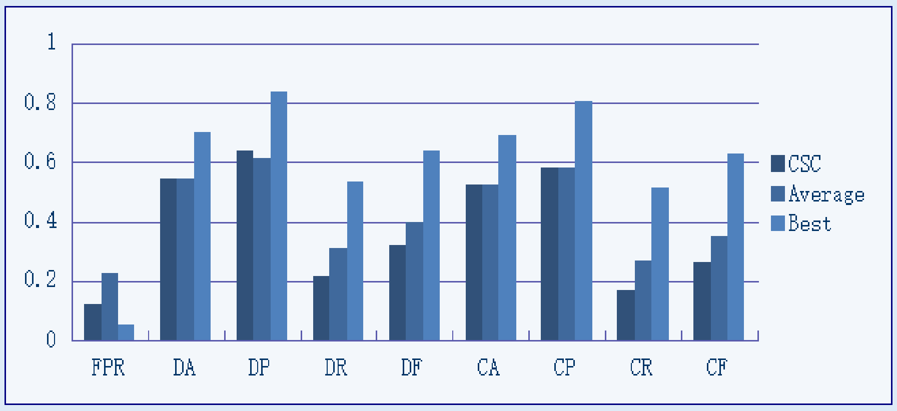


Figure 2. The Performance of the CSC system.

Introduction

Spelling check is a common task in every written language, which is an automatic mechanism to detect and correct human spelling errors.

Language modeling (LM) is widely used in CSC, and the most widely-used and well-practiced language model, by far, is the n-gram LM, because of its simplicity and fair predictive power. Continue to use N-gram LM, this paper proposed a model based on joint bi-gram and tri-gram LM to detect and correct spelling errors. And we try to exploit word segmentation in a pre-processing stage which improves the system performance to a certain extent. In addition, dynamic programming is applied to reduce the running time of our program and additive smoothing is used to solve the data sparseness problem in training set.

The Proposed System

Figure 1 shows the flowchart of our CSC system.

The system is mainly consists of four parts: Chinese Word Segmentation, Confusion sets, Corpus and Language Model. It performs CSC in the following steps:

Step 1. A given sentence was segmented by CSC system with Chinese words segmentation techniques. Result of Chinese words for segmentation will serve as the basis for the next step.

Step 2. According to the judgment conditions our system finds confusion sets of the corresponding word in the sentence.

Step 3. For each character in this sentence which can be replaced (in accordance with corresponding conditions), the system will enumerate every character of its confusion set to replace the original character. We will get a candidate sentence set after this step.

Step 4. The system will calculate the score of every candidate sentence by using the joint bigram and tri-gram LM (using bi-gram and trigram based on different conditions). We use the corpus of CCL and SOGOU to generate the frequency of n-gram. Finally, the sentence with the highest score will be chosen as the final output.

In order to decrease the running time in Step 3 and Step 4, we apply dynamic programming to optimize the algorithm.

Future Work

Figure 2 shows the performance of our CSC system.

It is our second attempt on Chinese spelling check, and the evaluation results of SIGHAN-8 CSC final test shows that comparing to the method we proposed in the CSC task of CLP-SIGHAN Bake-Off 2014 last year, we achieve a improvement of 9.7% in DF and 6.3% in CF. However, we still have a long way from the state-of-arts results. There are many possible and promising research directions for the near future. Language modeling has been extensively used in our CSC. However, the N-gram language models only aim at capturing the local contextual information or the lexical regularity of a language. Future work will explore long-span semantic information for language modeling to further improve the CSC. What's more, we still need to do more research on how to deal with the characters overkill problem to make the CSC more perfect.

Selected References:

- Shih-Hung Wu, Chao-Lin Liu, and Lung-Hao Lee. 2013. Chinese Spelling Check Evaluation at SIGHAN Bake-off 2013. In Proceedings of the 7th SIGHAN Workshop on Chinese Language
- Processing (SIGHAN'13), Nagoya, Japan, 14 October, 2013, pp. 35-42.
- Frederick Jelinek. 1999. Statistical Methods for Speech Recognition. The MIT Press.