An Evaluation of a Method to Detect and Correct Erroneous Characters in Japanese input through an OCR using Markov Models

Tetsuo Araki
Fukui University
3-9-1 Bunkyo,
Fukui-shi,Japan

Satoru Ikehara
NTT Communication
Science Laboratories
1-2356 Take,
Yokosuka-shi, Japan

Nobuyuki Tsukahara
Fukui University
3-9-1 Bunkyo,
Fukui-shi, Japan

Yasunori Komatsu
Fukui University
3-9-1 Bunkyo,
Fukui-shi, Japan

Abstract

The "*Selective Error Correction Method*" to judge these three types of errors, and correct them, using m-th order Markov chain model for Japanese 'kanji-kana' characters, has been proposed and shown to be useful to detect and correct errors generated randomly (Araki et al., 1994).

In this paper, this method is applied to detect and correct erroneous characters in Japanese text input through an OCR. The method is confirmed to be also effective to detect and correct the errors introduced by the OCR.

1 Introduction

In order to improve the computers' man-machine interfaces, input devices such as Optical Character Readers(OCR) or speech recognition devices have been developed. However, text input through an OCR or a speech recognition device usually contains erroneous character strings.

The erroneous characters can be classified into three types. The first is characters that have been recognized incorrectly , that is taken to be characters other than the correct characters. The second and the third are extra characters wrongly inserted and deleted (skipped) characters. Markov chain models have been used to find and correct the first type of errors.

Recently, the *Selective Error Correction Method* to judge the three types of the errors and correct correct these errors, using m-th order Markov chain model for Japanese 'kanji-kana' characters, has been proposed (Araki et al., 1994).

In this paper, the *Selective Error Correction Method* is applied to detect and correct erroneous characters in Japanese text input through an OCR.

2 Experimental Procedure and Conditions

An experimental procedure using the *Selective Error Correction Method* for erroneous Japanese phrases input through OCR is described in Fig.1.

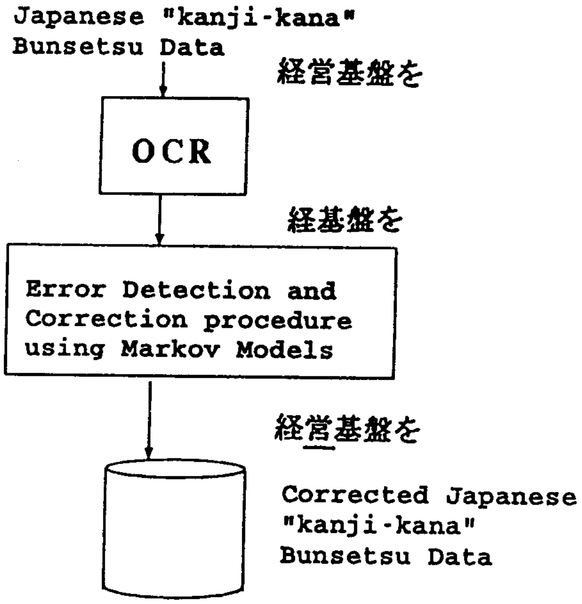

Fig.1. An Experimental Procedure Using Japanese Phrase Input Through OCR.

Experimental Conditions are denoted as follows:
(1) The number of phrases used for statistics: 70 issues of a daily Japanese newspaper containing 283,963 phrases.
(2) The number of phrases input through the OCR: 1000 phrases
(a) The average length of phrase (in 'kanji-kana' characters): 6 (b) The size of character fonts: 8 point
(c) The input method to the OCR: Fax

3 Basic Definitions and the Selective Error Correction Method using 2nd-Order Markov Model

A Japanese sentence can be separated into syntactic units called phrases (usually called "bunsetsu").

Japanese phrases in a text can be divided into two types: *correct* phrases, *erroneous* phrases. The set of *correct* Japanese phrases is represented by Γ_C. The set of *erroneous* phrases is denoted by Γ_E, and it is further divided into three types: The first is erroneous phrases which contain erroneous characters substituted wrongly in the phrase, and is denoted by Γ_S. The second and the third are erroneous phrases which have characters ommitted from them (denoted by Γ_D) or inserted wrongly in them (denoted by Γ_I).

The accuracy ratios to detect and to correct the errors by a method are evaluated by the "Relevance Factor" P and the "Recall Factor" R. Here, P denotes the ratio of errors detected or corrected by a method to the whole of Γ_E. R denotes the ratio that the elements of Γ_E can be detected or corrected by a method.

Next, we introduce the following assumption based on previous experiments: *"Each Markov probability for erroneous chains of syllables and 'kanji-kana' characters is small compared to that of correct chains"*.

According to this assumption, the procedure of detecting the location i and the length k of error chains is defined as followed: Namely, the procedure detects that k characters are wrongly substituted or inserted at the location i, if the m-th order Markov probability for a chain remains smaller than the critical value T just $(k + m)$ times from the location i to $i + k + m - 1$.

Similary, the method of detecting the location of a chain wrongly deleted in $\Gamma_D^{(k)}$ and the methods of correcting the chains with wrongly substituted, inserted or deleted characters are described in Ref.(Araki et al., 1994).

4 Experimental Results Using Erroneous Japanese Phrase Input Through OCR

4.1 Experimental Results

The critical value of the 2nd-order Markov probability T was determined so as to make the value of $P \times R$ maximum for *erroneous* phrases. The experimental results are described as follows:

[1] Error detection and error correction of correct phrases

All of *correct* phrase are judged to be correct.

[2] The Relation of P and R for *erroneous* phrases

The maximum values of P and R for the location of erroneous 'kanji-kana' character strings using error detection procedures and those of the errors corrected using error correction procedures, are as follows: (1) $P_S^{(D)} = 79.0\%$ $R_S^{(D)} = 74.5\%$ (2) $P_S^{(C)} = 66.2\%$ $R_S^{(C)} = 84.6\%$

The values of $R_S^{(D)}$ and $P_S^{(D)}$ mean that this method can find 74.5% of the *erroneous* phrases Γ_S (substitution type), and 21.0% of the errors detected by this method are errors detected wrongly.

From these results, it is shown that the *Selective Error Correction Method* using 2nd-order Markov models is useful to detect and correct erroneous characters substituted wrongly in text input through an OCR.

4.2 Discussion

[1] The characteristics of Erroneous Strings input through OCR.

Compared to the errors randomly generated (Araki et al., 1994), the errors caused by OCR showed high occurrence in the following four types of errors: (1) mixed type (combination of three error types), (2) errors located at the head and at end of phrases, (3) errors that length of an erroneous string in a phrase is greater than 3, and (4) errors distributed within a phrase.

[2] The comparison of the value of P and R for error detection and error correction.

The maximum values of P and R to detect and correct errors caused by an OCR are inferior to that of errors generated randomly by 20–40%.

The main reasons why the maximum values of P and R are reduced can mainly be explained by the characteristics of (2) and (4) above mentioned.

However, regarding to (1) substitution errors, (2) errors located inside phrases, (3) errors of length 1 and (4) errors connected in phrases, it is seen that the maximum values of P and R to detect and correct errors by OCR, are nearly equal to those for errors generated randomly.

5 Conclusion

In this paper, the *Selective Error Correction Method* proposed recently, is applied to detect and correct erroneous characters wrongly substituted, deleted and inserted in Japanese text input using an OCR, the method is shown to be effective, though the accuracy ratios to detect and correct the OCR errors is inferior to those of random errors.

References

T.Araki, S.Ikehara and N.Tukahara. 1994. An Evaluation to Detect and Correct Erroneous Characters Wrongly Substituted, Deleted and Inserted Japanese and English Sentences Using Markov Models. *COLING 94*, Vol.1,pp187-193.