LINGUISTIC ERROR CORRECTION OF JAPANESE SENTENCES

Tsutomu Kawada, Shin-ya Amano and Kunio Sakai

Information Systems Lab., Research and Development Center, Toshiba Corporation
1 Komukai Toshiba-cho, Saiwai-ku, Kawasaki 210, JAPAN

Abstract

This paper describes a newly developed linguistic error correction system, which can correct errors and rejections of Japanese sentences by using linguistic knowledge.

Conventional optical character readers (OCR) need human assistance to correct their recognition errors and rejections. An operator must teach the OCR correct answers whenever an illegible character pattern occurs. If this error correction operation is mechanized, the throughput of the OCR will increase.

This linguistic error correction system offers means of automated error correction by analysing sentences of the OCR outputs linguistically. This system grammatically selects legal letters from the candidates which can not be decided uniquely by pattern recognition only, and recommends grammatically and semantically meaningful letters for the illegible letter.

1. Introduction

More than 2,000 different chinese characters are used in Japanese newspapers and publications. The large repertory of letters as well as the structual complexity of each character pattern are the main difficulties to recognize letters by OCR. Mutually similar character patterns mainly cause recognition errors and rejections. The difference between such similar letters usually concentrate on a local area [滅 ···· 滅 , 徴 ···· 徴 , 億 ···· 億 , ハ ···· ハ , 大 ···· 犬 ···· 太 , etc]. The output of OCR contains these ambiguities.

It is necessary to use the contextual information contained in every kind of natural text. The application of context makes it possible to detect errors or even to correct them. The main object of this error correction system is to resolve these ambiguities on the basis of linguistic knowledge. To resolve these ambiguities each ambiguous letter is put in its sentence, and the sentence is analyzed syntactically and semantically. When the sentences is acceptable the letter is selected preferably.

There has been some contextual post processing systems.[1~2] But these systems are designed to read only postal-addresses. The words which are included in the postal-address are restricted within a relatively narrow domain. The postal address usually consists of a person's name, place name, postal code and numbers. On the other hand this newly developed kanji OCR can read actual texts of Japanese sentences, and its contextual error correction system of it has been designed to deal with many kinds of words of various parts of speech (noun, verb, adjective, adverb, pronoun, etc.). This error correction system has a practical 25,000 words dictionary, and can correct 53.8 percent of the errors which are included in the outputs of kanji OCR.

2. Restrictions

This system imposes two restrictions on input data. One is that the input must consist of grammatical Japanese sentences in order that syntax analysis can be applicable. This system is not effective for only numeral data texts or a mere list of words.

The other restriction is that the texts to be dealt with must be limited to a special field. By this restriction we can limit the number of terminologies which are used in the field.

The corpus used for the experiment is 1,700 claims of patent gazettes of the Japan Patent Office. These gazettes concern the manufacturing technology of LSI devices for thirteen years (1964 - 1976). Figure 1 shows an example of them. This corpus includes 306,000 words. There are about 5 thousand different words in it. The distribution of the various categories are as follows;

noun	3603	functional word	90
verb	832	suffix and prefix	110
adjective	75		
adverb	61		
conjunction	30		

As twenty thousand common words are added to these words, the dictionary contains about twenty-five thousand words.

3. Features of Patent Sentences

A Japanese written language consists of kanji, kana (hirakana and katakana) and alphanumeric letters. Kana is a phonetic symbol and kanji is an ideograph. The kana set (either hiragana on katakana) consists of 48 letters. More than 2,000 different kanji letters are daily used.

Japanese people write a sentence like one

㈭ＭＯＳ型半導体集積回路装置

㉑特　　願　昭４８－３３６４
㉒出　　願　昭４７（１９７２）１２月２８日
　公　　開　昭４９－９０８９２
　　㊸昭４９（１９７４）８月３０日
㉒発　明　者　大曽根雄志
　　門真市大字門真１００６松下電器
　　産業株式会社内
　同　　平尾孝
　　同所
㉑出　願　人　松下電器産業株式会社
　　門真市大字門真１００６
㉔代　理　人　弁理士　中尾敏男　外１名

bibliography

㊼特許請求の範囲

１　半導体基板内に形成されたソース、ドレイン拡散領域およびこれらの領域間に設置されたゲート酸化膜、ゲート電極よりなるＭＯＳトランジスタと、このＭＯＳトランジスタ上に形成された高比抵抗の多結晶シリコン層と、上記多結晶シリコン層の一部に不純物を注入して形成されるとともに上記ＭＯＳトランジスタの負荷抵抗となる抵抗体とを備えたことを特徴とするＭＯＳ型半導体集積回路装置。

claim sentence

Figure 1. Actual Patent Gazette

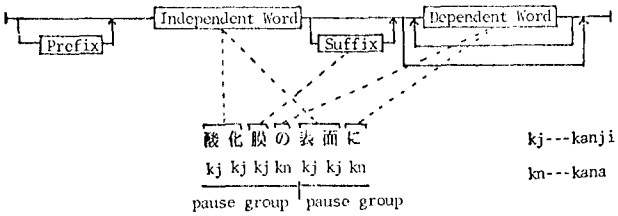

Figure 2. Construction of a Pause Group

continuous string of letters with no space (see Fig. 1). Japanese is different from western languages in this point. It is firstly important to identify words in the continuous string of letters to analyze a Japanese sentence. Figure 2 shows the construction of a pause group which is the minimum meaningful unit of a Japanese sentence. The prefix, the independent word and the suffix are usually written in kanji or katakana letters. The dependent word is written in kana letters. Changes of letter types as well as punctuation symbols give us useful clues to the boundaries where it is possible to separate a long letter string into shorter manageable units (pause group). This correction system detects words by using such conditions for these boundaries (Fig. 2).

Experiments were conducted for the claim sentences of the patents. A claim sentence has a particular style. Most of the claim sentences consist of one sentence. An analytical study[3] of the claim sentences showed that all sentences were categorized into 14 sentence patterns by coordinate phrases. The average count of words for a sentence is 180 words. The sentence is a big noun phrase and is constructed from many coordinate adjective or adverbial phrases which modify the same word. The claim sentence is so long that it is practical to analyze it on the basis of these phrases.

4. Kanji OCR and it's Errors

The large number of character categories as well as structual complexity of each character pattern are the dominant difficulties in kanji character recognition. A two-stage recognition method[4] has been developed to cope with these difficulties. This method employs an efficient candidate selection prior to a precise individual recognition. Fig. 3 shows a diagram of the two-stage recognition method.

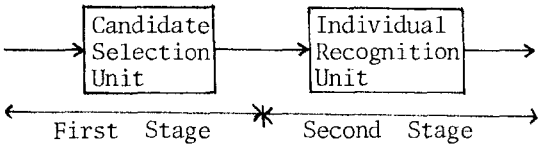

Figure 3. Two-Stage Recognition Method Diagram

In the first recognition process stage, feature extraction is carried out on the input pattern. Candidate characters are obtained according to their geometrical features. In the second stage, pattern matchings are carried out between the input pattern and each reference pattern in selected candidate characters. The decision is made on the basis of their similarity values.

The mutually similar patterns as well as low print quality cause recognition errors and rejections. These illegible letters have low similarity values. The recognition speed of this kanji OCR is 100 characters per second. More than 99 percent correct recognition rate was obtained for actual data. The average letter count of the claim sentences is 450 letters. Consequently this system encounters an illegible letter every second and three or four letters in a claim sentence.

As the illegible letters have low similar-

ity values, this correction system can find doubtful letters easily. If this error correction system checks all letters which are contained in a text, it needs much time to process. This error correction system picks out only the phrases which contain illegible letters, and analyze the grammatical legality of them. By this restriction this error correction system decreases the processing time and becomes a practical one.

5. Error Correction Method

The error correction system has three analysis functions (Fig. 4).
a) word analysis function
b) syntax analysis function
c) wording analysis function

Two notations are used here. When one letter can not be recognized uniquely, the candidates for the letter are enclosed in parentheses. A letter which can not be recognized at all is expressed by a question mark.

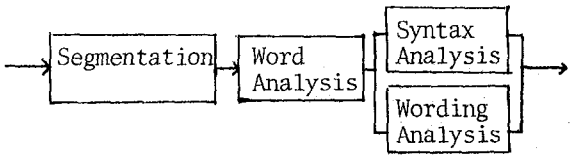

Figure 4. Error Correction Method Diagram

5.1 Word Analysis Function

In case of encountering an ambiguous letter in a sentence, the word analysis program searches the dictionary to find a grammatically and semantically valid candidate.

For example, '[パ バ]ターン認識 ([PA BA]TANNINSIKI)' shows that the first letter is an ambiguous letter. In this case, two candidates are tested. A candidate 'パターン認識 (pattern recognition)' is meaningful but 'バターン認識' is not. So パ is determined as the unique answer.

Some Japanese letters resemble closely.

Example 1.

	Letter	Meaning
ベ[----]ス	ー	× minus
	―	× dash
	一	× one
	—	○ symbol for long vowel

The selection from these resembling patterns depends on their context. '- (a long vowel)' is reasonable for ベ[----]ス (BEESU, base).

Two or more words are frequently connected without any conjunction or preposition in Japanese sentences. In this case the word analysis program calls the compound word analysis subprogram which looks up the word dictionary and makes a compound word from two or three words to analyze it. The above example is a compound word. 'パターン認識 (pattern Recognition)' is a compound word constructed from 'パターン (pattern)' and '認識 (recognition)'. This subprogram has not only a full-string but a sub-string matching ability (Fig. 5).

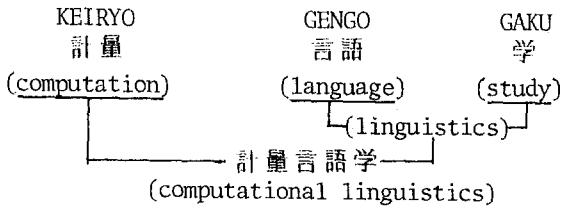

Figure 5. Sub-string Matching of a Compound Word

When a letter has no candidate letter, the word analysis program consults the dictionary and searches for words which fill up the illegible letter.

Example 2.

?明	発明	inovation	noun,verb
	証明	proof	noun,verb
	表明	manifestation	noun,verb
	文明	civilization	noun
	照明	lighting	noun,verb
	透明	transparency	adjective verb
	判明	become clear	noun,verb

In this case seven letters fill up the ?. The selection from these candidates is performed in the next syntax analysis step. This word analysis program is not valid for consecutive illegible letters. As most of the Japanese words are one or two Kanji letters, consecutive illegible letters do not give us any clue to search the dictionary. When we are given consecutive illegible letters '??', we can hardly guess what they are.

5.2 Syntax Analysis Function

When the word analysis is unsatisfactory to resolve the ambiguities, the syntax analysis is applied to them. In example 2, there are still seven candidates which were selected by the word analysis function. The syntax analysis program refers the contextual information.

Example 3. ?明な電極 (transparent terminal)

?明な電極	発、証、表、文、照、判	×
	透	○

This program first conducts a morphological analysis of the given pause group, and analyzes the syntactic role of each pause group in its phrase. A noun or verb does not conjugate like ' な(NA), and only an adjective verb can conjugate like ' な(NA)'. So ' 透明 ' is selected uniquely.

Example 4. 基板［が、か］ある KABAN(GA KA)ARU
基板がある There is a base.
基板かある

In this case ' 基板 (KIBAN; base)' in a noun, ' ある (ARU; be)' is a verb, ' が (GA)' is a particle to indicate the subject, and ' か (KA)' is a particle to indicate an alternative or question. The particle ' が ' only makes the sentence grammatical and ' 基板がある ' is the unique answer.

This syntax analysis program performs the morphological analysis to the segmented pause groups (Fig. 2). If the segmentation is incorrect, this program can not analyze the phrase or sentence. So this program retries the segmentation of the input string to make successful analysis results.

Example 5. a) 制御回路に［そぞ］れぞれ加え

b) 制御回路に［そぞ］れぞれ／加え

c) 制御回路に／それぞれ／加え
(control circuit) (each) (add)

This example shows the retry process. The segmentation program firstly segments a string at the point of letter type changing (b). This segmentation is not correct. The first pause group is not grammatical. This program assumes that this pause group may be a compound pause group, and searches all possible separations from left to right. This program finds a embedded adverb ' それぞれ (SOREZORE; each)', and by this segmentation this sentence can be analyzed successfully. The other candidate ' ぞ ' can make no grammatical sentence.

5.3 Wording Analysis Function

In the sentences of patent gazettes, important words or key words are repeatedly used with anapholic pronouns. This fact is a very important clue to find an anaphola or to guess the ambiguous letter. The arrows in Fig. 6 show the anapholic relations of words in a text. Some kinds of particular anapholic pronouns appear in patent texts (' 上記 (above-mentioned)', ' 該 (GAI; such)', ' 同 (DOU; same)' and ' この (KONO; this)'). When an illegible letter occurs in an anapholic words, the wording analysis program searches the indicated word and correct the illegible letter by the matched letter. In Fig. 6, ' 上記接合？域

特許請求の範囲

1 導電型半導体基板内に形成された反射導電型の接合領域と、この接合領域に隣接して上記基板内に形成された上記基板と同一導電型の高不純物濃度領域と、上記接合？域と高不純物濃度領域が互いに重なり合う部分の上部に絶縁膜を介して形成されたゲート電極とを備えたことを特徴とするMOS型ダイオード。

Figure 6. Anapholic Relations in a Sentence

(above-mentioned connected area)' has an anapholic pronoun ' 上記 '. ' 接合？域 ' is compared with the indicated word ' 接合領域 ', and ? is corrected to ' 領 '. The wording analysis program automatically prints out a glossary of texts. This glossary is used to augment the dictionary of the error correction system.

Numeric expressions are also used frequently. Numeric expressions are analyzed by using semantic relations of words in their vicinity. As the bibliography of a patent contains the name of a person, place and affiliated organization, the correction system needs to change the dictionary from a common dictionary to a proper noun dictionary. In a proper noun pause group, it is more important to analyze the semantic relation among the words.[5]

Example 6.

神奈川県 ？崎市 (KAWASAKI city KANAGAWA prefecture)
KANAGAWA KEN KAWASAKI SHI
川崎 市 (name of city)
河崎 氏 (person's name)

This phrase describes an address, and ' 市 (city)' is a suffix for the name of a place which does not connect with a person's name. So the illegible letter can be decided uniquely.

6. System Configuration and Experimental Results

Fig. 7 shows the kanji OCR and linguistic error correction system. Fig. 8 shows the configuration of this system. The error correction system is programmed on a mini-computer (TOSBAC-40). The text editing terminal is a newly developed Japanese word processor. The operator of this system can confirm the error correction results on the CRT display, change the form of the text by versatile editing functions, store and transfer them to the host machine.

The experimental results for actual 250

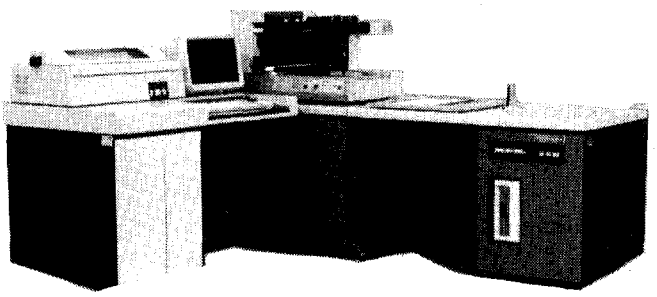

Figure 7. Overall View of Error Correction System

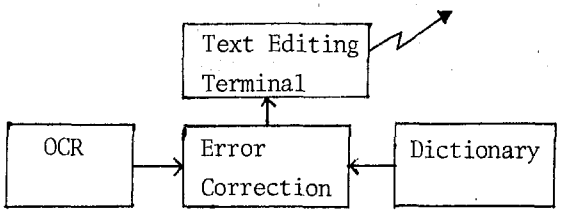

Figure 8. System Configuration

patent texts were as follows;

effective correction ..... 53.8 percent
ineffective correction ..... 38.5 percent
wrong correction ..... 7.7 percent

The ineffective correction rate shows the percentage of letters which this system can not correct.

Example 7. Wrong correction

？件請求の… (What we claim is---)

[用準条本物]件請求の -------wrong
(YOU JI JYOU HON BUTU)KEN SEI KYU NO

特許請求の ----------right

This example shows a case of wrong correction. The first letter was illegible. And the next letter '件 (KEN)' was misread. The correct letter is '許 (KYO)'. The kanji OCR has made an error. This error correction system tried to correct the ? letter by using '件' which was a wrong letter as the clue for correction, and made a wrong correction.

7. Conclusion

This error correction system can correct about fifty percent of the errors and rejections of kanji OCR outputs and was effective to increase the total throughput of the kanji OCR.

The kanji OCR reads letters according to their geometrical feature, and this linguistic error correction system reads a sentence according to the linguistic knowledge. The combination of the kanji OCR and linguistic error correction system realizes a practical Japanese text reader and can cope with the increasing demands for input of Japanese document information. The throughput rate of the OCR, combined with this linguistic error correction system, is about 10 times higher than that of a conventional manual data entry.

8. Acknowledgments

Parts of the research, and development of the system were made under contract with the Ministry of International Trade and Industry on the Pattern Information Processing System (PIPS) Project.

9. References

(1) S. Viresh, "An Approach to Address Identification from Degraded Address Data", Proc. NCC pp.779-783, 1977.

(2) E. M. Riseman, A. R. Hanson, "A contextual Postprocessing System for Error Correction Using Binary n-Grams", Trans. on COMPUTER IEEE, Vol. C-23, No.5, MAY, pp.480-493, 1974.

(3) H. Saiito, M. Noyori, "Patterns of Claim of Japanese Patent Sentences" computational linguistics, IPSJ, pp.1-10, Feb. 1978.

(4) K. Sakai, S. Hirai, T. Kawada and S. Amano, "An Optical Chinese Character Reader", Proc. Third IJCPR, pp.122-126, 1976.

(5) T. Kawada, S. Amano, K. Mori and K. Kodama, "Japanese Word Processor JW-10", Proc. COMPCON'79, pp.238-242, Sept. 1979.