MVA '92 IAPR Workshop on Machine Vision Applications Dec. 7-9,1992, Tokyo

IMPROVEMENT OF TEXT IMAGE RECOGNITION BASED ON LINGUISTIC CONSTRAINTS

Koichi KISE, Tadamichi SHIRAISHI, Shinobu TAKAMATSU and Hiroji KUSAKA

Dept. of Electrical Eng. Univ. of Osaka Prefecture
1-1 Gakuencho, Sakai, Osaka 593 Japan
E-mail: kise@neptune.denki.osakafu-u.ac.jp

Abstract

This paper describes a new method of post-processing for character recognition, which utilizes a natural language processing (NLP) system. Our method is based on a word matching method which verify the results of character recognition by matching them with dictionary words. The characteristics of our method are as follows: 1)The word matching process is guided based on the syntactic and semantic constraints on sentences to improve the efficiency, 2)Top-down assumption of characters by the NLP system enables us to correct errors which are difficult to be recovered by the word matching, 3)The facility for re-recognition of characters is introduced to determine which extracted word is most plausible in case there is little evidence derived from syntactic and semantic constraints. From the experiments for 50 images of sentences, we demonstrate the effectiveness of our method.

1 Introduction

Text image recognition is the process to extract the information recorded as characters in text images. In pursuit of robustness and accuracy of the process, it is indispensable to cope with errors in character recognition caused by the imaging defects, variety of font styles, characters having similar shapes, etc. Since improvement of a method of character recognition alone has its limits, an advisable solution is to introduce the post-processing of character recognition.

From this point of view, numerous attempts have been made over the past few decades [1]. A word matching method, which verifies a sequence of recognized characters by an approximate matching with dictionary words, is one of the promising ways to correct the errors. The approximate word matching is a method of string matching which allows some disagreement characters in a word.

For images of severely low quality, however, we face the problem to determine the most plausible word from the words extracted by the word matching; In general, it is necessary to allow a lot of disagreement characters in the word matching to cope with the frequent errors. This results in accepting many erroneous words.

The problem is more serious for the recognition of Japanese texts. Texts written in Japanese do not include word separators such as spaces in English, thus it is also required for the word matching to segment words from a sequence of characters. As a result, enormous number of possible words would be extracted by the word matching.

One expedient way to select the most plausible word is to utilize a similarity score obtained by the character recognition. In this case, however, it does not contain fruitful information due to the low quality of images. An alternative way is to exploit *inter-word* constraints to reduce the number of possible words. Some researchers attempt to use probabilistic or morphological constraints between words[2, 3].

In this paper, we describe a method of post-processing which exploits the constraints on syntax and semantics of sentences with the help of a natural language processing (NLP) system. The characteristic of our method is summarized to the point that these constraints are utilized not only for the verification of extracted words, but also for the extraction of words: An extracted word is verified whether it is consistent with the previously extracted words in the sense of syntax and semantics, so as to find an inconsistent word as soon as possible. In addition, the dictionary words used for the matching are reduced based on the syntactic constraints obtained from the previously extracted words.

The effectiveness of our method is demonstrated by the experiments for 50 images of sample sentences in comparison with the case that a word matching method is solely applied.

2 Text Image Recognition System

Figure 1 illustrates the architecture of our text image recognition system. Our system consists of three modules: pre-processing module, character recognition module and post-processing module.

The task for the pre-processing module includes noise reduction, skew correction, and segmentation of textlines and characters. In the current implementation, we assume that no error occurs in the segmentation.

After the pre-processing, the segmented character images are recognized by the character recognition module. To cope with the difficulty of recognition, multiple candidate characters are allowed for each character region. As shown in Fig. 2, these candidates are listed in descending order of shape similarity scores obtained through the recognition. In this paper, this is called a list of candidate characters. The number of candidates is currently fixed to 5 for each character.

The task for post-processing can be regarded as traversing the search space spanned by the list of candidate characters to select correct candidates. The post-processing module utilizes two sub-modules for this purpose: the word matching unit and the NLP system. The word matching unit is controlled by the NLP system to extract a sequence of words which is valid as a sentence from the viewpoint of syntax and semantics. In addition, the character recognition module is utilized again

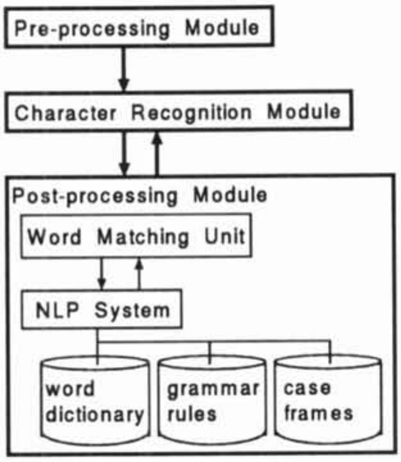

Figure 1: System architecture

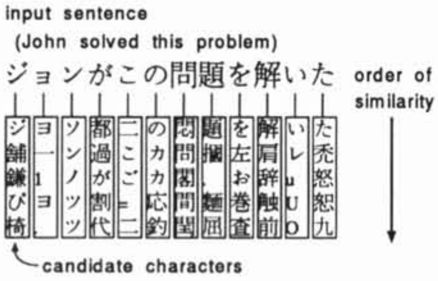

Figure 2: A list of candidate characters

to determine which extracted words is most plausible in case there is little evidence to select them by the post-processing module.

3 Natural Language Processing

In this section, we will show the brief overview of our NLP method which analyzes syntax and semantics of a given sentence. This method is originally developed for the task of machine translation. For further details, see[4].

3.1 Syntactic analysis

Our method of syntactic analysis is classified into the left corner branch method based on the reachability matrix. The NLP system analyzes a given sentence from its left to the right (i.e. the reading order of a sentence) in a bottom-up manner. Context free grammar rules (simply called grammar rules in this paper) are employed for the analysis. Top-down information derived from the reachability matrix is also utilized to avoid the futile application of grammar rules.

From the viewpoint of the post-processing, the advantage of our NLP system can be thought of as its facility to predict a part of speech; a part of speech for the next word is predicted based on the previously analyzed sequence of words. This enables us to reduce the number of dictionary words tested by the word matching unit, so that the efficiency of the post-processing is improved.

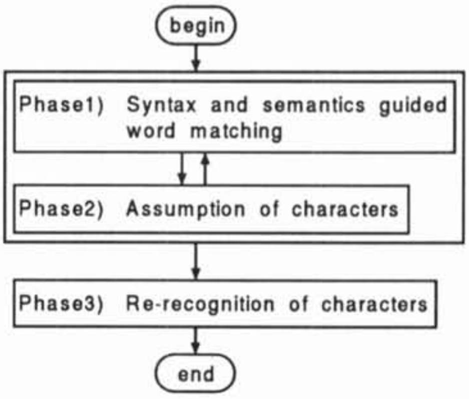
begin
Phase1) Syntax and semantics guided word matching
Phase2) Assumption of characters
Phase3) Re-recognition of characters
end

Figure 3: Control flow of the post-processing

3.2 Semantic analysis

The method of semantic analysis incorporated in our NLP system is based on *semantic categories* of words and *case frames* for verbs. A semantic category represents a label for the meaning of a word such as "physical object" and "mental object". A semantic category of each word is recorded in the word dictionary. On the other hand, a case frame describes the legal patterns of case structure for a verb in terms of case labels for words such as "agent" and "object", as well as their semantic categories.

In our NLP system, semantics of a sentence is verified by matching between case frames and semantic categories of words. This is executed by the additional terms of the grammar rules. These terms work as the condition for the application of a grammar rule, thus semantic analysis is simultaneously performed with the syntactic one.

4 Post-processing

4.1 Overview

Figure 3 shows the control flow of the post-processing consisting of 3 phases. The input is the list of candidate characters, while the output is a sentence composed of the extracted words.

The post-processing module normally performs the phase 1 to find a sequence of words in a given sentence. In the phase 1, approximate word matching is employed to correct recognition errors.

In case the phase 1 fails, the post-processing module tries the phase 2 to assume missing characters in the list of candidate characters on condition that the syntactic or semantic constraints for the next word are strong enough. After the phase 2, the phase 1 is resumed to extract the following words. In case the phase 2 is impossible to be applied, backtracking occurs in the phase 1 to retry the word matching.

The phase 3 is executed when the word matching finishes. The task of this phase is to solve the ambiguity in the extracted sequence of words. In case there exists multiple words whose positions in a sentence overlap with each other, the characters included in the words are recognized again to determine which is the most plausible word in terms of its shape.

In the following sections, we will describe each phase in detail.

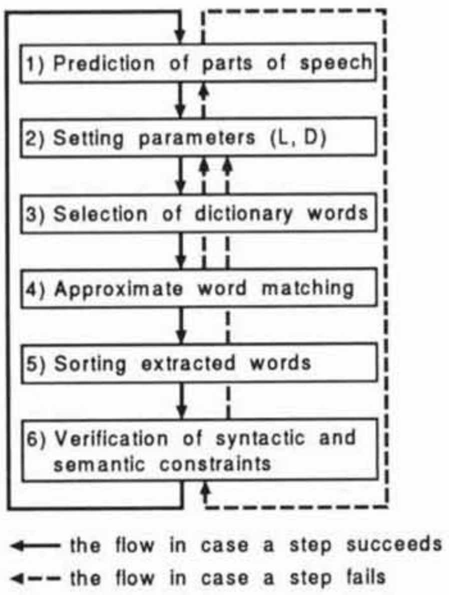
1) Prediction of parts of speech
2) Setting parameters (L, D)
3) Selection of dictionary words
4) Approximate word matching
5) Sorting extracted words
6) Verification of syntactic and semantic constraints
the flow in case a step succeeds
the flow in case a step fails

Figure 4: Algorithm of the word matching

4.2 Syntax and semantics guided word matching

The algorithm of the phase 1 is shown in Fig. 4

1) Prediction of parts of speech At the first step, the NLP system predicts possible parts of speech for a word in a top-down manner. This prediction is based on the already extracted sequence of words, the reachability matrix deduced from the grammar rules. In case multiple parts of speech are predicted, one of them are randomly selected; the rests are retained for the later examination. The part of speech selected in this step is determined to be inappropriate in case the step2 fails. If all possible parts of speech are found to be inappropriate, this step fails.

2) Setting parameter The next step is to set parameters to control the following steps. We use the following two parameters L and D, for retrieving dictionary words (step3) and for the word matching (step4), respectively.

- L is the length of a dictionary word to be retrieved, (i.e. the number of characters contained in a word)
- D is the upper limit of the rate of disagreement characters for the length of a dictionary word.

In the current implementation, D is restricted to one of the discrete values of 0%, 30%, 50%.

At the first trial of the word matching, $D = 0\%$ and the largest value of L are selected. This indicates that we prefer exact matching and the longest word. In case the word matching (step4) fails under this condition, the values of D and L are changed for the next trial. First, a smaller value of L is applied with the current value of D. If the matching fails with all possible values of L, then D is loosened. In case the new value of D is selected, L is reset to the largest value, and then decreased. If the following steps fail under all possible conditions, this step fails.

3) Selection of dictionary words At the third step, dictionary words are retrieved based on the predicted part of speech and the word length L. In general, multiple words are retrieved from the dictionary. All of these words are reported to the word matching unit.

4) Approximate word matching At the fourth step, matching between each dictionary word and the list of candidate characters is performed to extract all possible words. To cope with the recognition errors (i.e. the case that the correct character is missed in the list of candidates), we utilize the approximate matching specified by D mentioned above. For example, under the condition of $D = 50\%$ and the word length $L = 3$, one disagreement character is allowed in the word matching. If no word can be extracted, this step fails.

5) Sorting extracted words The fifth step is for sorting the extracted words to record the words in a stack. The score for sorting is the sum of the orders of the candidate characters included in a extracted word. A word with a smaller score holds upper position in the stack. Note that a disagreement character can be neglected to calculate the score, because the number of disagreement characters is fixed in advance by the value of D. In case multiple words have the same score, they are sorted randomly.

6) Verification of syntactic and semantic constraints At the last step, the word is popped from the stack to verify whether it is consistent with the sequence of previously extracted words in the sense of syntax and semantics. If it is inconsistent, the next word is popped. Otherwise, it is assumed to be correct in order to extract the succeeding words. In case the stack is empty, this step fails.

4.3 Assumption of characters

In most cases, the method of approximate word matching utilized in the phase 1 is powerful enough to correct errors of character recognition. However, it is impossible to correct words which consist of one character (one-character word). This causes a serious problem, because there exists many one-character words in Japanese sentences. For example, most of the case particles are one-character words.

Fortunately, the syntactic and semantic constraints on case particles are strong enough to assume it, even if there is no candidate in the list. For example, suppose the case particle 'が' is missed in the list of candidates in Fig. 2. the phase 1 fails after extracting 'ジョン'(John). Based on the prediction of parts of speech, it is found that the next word would be a case particle. At the phase 2, the post-processing module assumes that the next word is a one-character word of a case particle to continue the word matching in phase 1.

After a verb in a sentence is successfully extracted by the phase 1, the case frame for the verb is available to determine which case particle is appropriate for the assumed character. For this example, the case particles 'が' and 'は' are selected.

In addition, an inflectional ending of a word can also be assumed based on the syntactic constraints.

4.4 Re-recognition of characters

There exists two cases a word cannot be determined by the phases mentioned above:

A) There remains words which have the same score.

B) Multiple words are assumed for the overlapping regions by the phase 2.

An example of the case B has already described in the previous section. Here we will show an example of the case A. Suppose the character ' こ ' is missed out of the list of candidate characters in Fig. 2. Three pronouns

'この'(this), 'あの'(that) and 'その'(it) are extracted by the approximate word matching with $D = 50\%$. Note that we have no information to determine which is the most plausible word, because all of the underlined characters are assumed by the phase 1. In such a case, these characters are recognized again for the selection based on their shapes.

5 Experimental Results

To verify the effectiveness of our method, experiments are conducted for images of 50 sample sentences including 637 characters. These sentences are picked up from a textbook of Japanese grammar, so that they have various types of syntactic and semantic patterns. For the experiment, we utilized the word dictionary including 412 words, 24 grammar rules, 50 semantic categories and 129 case frames.

For the purpose of comparison, the following two methods are applied besides our method:

A) Word matching method

B) Our method without the re-recognition phase

The method A extracts all possible words solely by the word matching which is the same one used in our method, except that the value of D is fixed to 50%. Output sentences are generated by combining the extracted words in such a way that no regions of words overlap with each other. The purpose of applying this method is to measure the potential size of the search space traversed by the word matching, as well as the accuracy of the word matching. On the other hand, the method B is applied to demonstrate the effectiveness of the syntactic and semantic constraints on sentences, by detaching the re-recognition phase.

The performance of these methods is measured by the following two criteria: the number of generated sentences (N), and the correct rate of generated sentences (R). The number of generated sentences (N) indicates that the average number of sentences generated for one sample image. The correct rate of generated sentences (R) is to measure the ability of generating a correct sentence on condition that multiple sentences are allowed for one sample. It is defined as $R = N_c/N_s$ where N_c and N_s denote the number of correct sentences included in the generated sentences, and the number of sample sentences, respectively.

The results are shown in Table 1. Note that the character recognition rate for sample sentences is 83.0% without the post-processing. The value of N for the method A demonstrates that the combinatorial explosion occurs in case the word matching is solely applied. In addition, the method A is not so accurate because the value of R is only 74.0%. This shows that only 74.0% of sentences can be recognized even if all correct sentences are selected from the output.

The results for the method B indicate that the syntactic and semantic constraints are effective enough to reduce the enormous search space. These constraints are also effective to correct errors in the character recognition, because R is notably improved in comparison with that for the method A.

Table 1: Experimental results

	N	R
Method A	4.5×10^5	74.0%
Method B	1.18	94.0%
Our method	1.0	94.0%

The results for our method represent that the correct sentences can be selected by the re-recognition phase without degrading the correct rate of generated sentences (R). Focusing on characters, the recognition rate of 83.0% is improved to 98.0%.

From these results, we can conclude that our method is effective and efficient enough to correct errors of character recognition in images of low quality.

6 Conclusion

We have presented a new method of post-processing, which incorporates NLP system to improve the effectiveness and efficiency of the approximate word matching. The characteristics of our method are summarized as follows:

- The dictionary words tested at each step of the word matching are limited based on the syntactic constraints.
- An extracted word is verified from the viewpoints of syntax and semantics.
- The facility for assumption of characters enables us to correct errors of character recognition in one-character words.
- Re-recognition of characters is employed to reduce the ambiguity of extracted words.

The future work to be explored is to cope with the errors of character segmentation.

References

[1] Elliman, D.G. and Lancaster, I.T.: "A Review of Segmentation and Contextual Analysis Techniques for Text Recognition", *Pattern Recognition*, Vol.23, No.3/4, pp.377–346(1990).

[2] Takano, T. and Nishino, F.: "Implementation and Evaluation of Post-processing for Japanese Document Readers", *Trans. IPSJ*, Vol.30, No.11, pp.1394–1401(1989) [In Japanese].

[3] Ikeda, K., Ohta, Y. and Ueno, E.: "Vocabular and Contextual Postprocessing for the Recognition of Hand-printed Japanese Manuscript", *Trans. IPSJ*, Vol.26, No.5, pp.862–869(1985) [In Japanese].

[4] Nishida, F. and Takamatsu, S.: "Automated Procedures for the Improvement of a Machine Translation System by Feedback from Postediting", *Machine Translation*, Vol.5, pp.223–246(1990).