An Interactive Rewriting Tool for Machine Acceptable Sentences

Hideki Hirakawa*, Kouichi Nomura**, Mariko Nakamura**

* Toshiba Corporation, R&D Center 1, Komukaitoshiba-cho, Saiwai-ku, Kawasaki 210, Japan hirakawa@isl.rdc.toshiba.co.jp

Abstract

This paper proposes an interactive rewriting tool for supporting people in creating machine acceptable sentences. The experiment and evaluation of this tool conducted when applied to a pre-editing support tool in a Japanese-to-English machine translation (henceforth, MT) system are also described.

1. Introduction

Supporting people in creating machine acceptable sentences is one important issue of future NLP applications. Agents in computers will be able to provide people with a variety of services by grasping the contents of documents. An intelligent mail server, for example, can read mails and send appropriate replies semi or fully automatically if the mails are machine acceptable. This is a natural extension of the ideas explored in the object lens(Malone, Grant, Lai, Rao and Rosenblitt, 1986).

However, it is difficult for humans to write in machine acceptable forms because of gaps between humans and machines. First, humans use semantics, context and common sense in interpreting a sentence, while limited resources are available to machines. Second, unlike machines, humans cannot simulate sentence analysis process in a machine to check all possibilities. As such, man-machine corporation would be one solution with the least burdens on humans.

Prior to a general-purpose rewriting tool, we have been developing a pre-editing support tool for an MT system(Hirakawa,Nogami and Amano,1991). Since our central concern is to analyze human-computer interaction for creating machine acceptable sentences, MT systems for bilingual and monolingual users are closely related to our tool. Considering that the difficulty of interactive systems lies in "so many interactions for a single sentence,"(Hutchins and Somers, 1992) we need a framework which produces the best result with least interactions.

2. Interactive Rewriting Tool

2.1 Design Principles

The problems of rewriting include difficulties in learning its know-how and unpredictable effects of rewriting. The following are our design principles:

1) <u>Interaction-based system</u>: Since rewriting is a dynamic process, the process should be interactive to deal with changes.

** Toshiba Corporation, Tokyo System Center 3-22, Kata-machi, Fuchuu, Tokyo 183, Japan {nomura, mari}@docs.sc.fuchu.toshiba.co.jp

2) <u>Presentation of rewriting candidates</u>: The system should present possible recommended revisions to users so that they can select the best choice and guarantee the correctness of the rewriting.

3) <u>Minimization of interactions</u>: To minimize frequencies of interactions in rewriting, the system should fully utilize knowledge available to improve the accuracy of the diagnoses.

4) <u>Optimization of interactions</u>: Scales to measure the degree of rewriting(scalability) should be introduced to optimize interactions to obtain the maximum effects with minimum interactions.

2.2 System Configuration

The system consists of two main parts: a sentence checker and a user interface unit

The sentence checker is composed of a sentence analyzer, an information extractor and a rewriting candidate generator. The information extractor extracts the information necessary for rewriting, such as morphological and syntactic information, from the analysis results of an MT system, to detect problematic phrases and generate recommended rewriting examples along with guidance messages to help users with rewriting.

The user interface displays the original sentences which require correction with their recommended

」 連用中止 」	文字 「」ひらかな	
(不要句) (並列句	リ 「挿入句」 辞書登	秋) 再解析)
	<u>) 削除</u>) <u>取消</u>) 文切文字)
tolnet接続を対象解了したい場合、標準エスケープ文字(通常CTRL+)すなわち CTRLキーを押し、同時に大かっこ"]"キーを押す)を フロンプトに対しguitを入力します。		
Input Sentence Expression which needs rewriting		
書表之候補: 連用中止		[] 検出項目名
入力することにより、 入力するので、 入力しなから、 入力します。 入力します。 入力します。そして、	Choices for rewriting	 幺 長文 幺 運用中止
入力します。そのため.		
理由、ポイント: 用言と用言との意味関係) き直してください。	Directions for rewriting	Items to be detected
メッセージ: 書換えが必要です。理由 この検出項目は複数の場		•

Fig. 1 User Interface Unit Screen

rewriting examples and guidance messages as in Fig. 1. When there are several problems in one sentence, they are presented to users in order of importance.

Knowledge for rewriting is accumulated as rules for the information extractor, which is actually a general-purpose information extracting tool-kit equipped with overall linguistic analyzing modules and information extracting/diagnosing modules.

3. Knowledge for Rewriting Long Sentences

Of those Japanese expressions that need rewriting before machine translation, long sentences which should be divided into shorter ones are the most important, as shown in previous studies(Kim, Ehara and Aizawa, 1992).

3.1 Criteria for Detecting Long Sentences

It is empirically known that simple factors, such as the number of characters or words in a sentence, are not sufficient to determine which sentences should be rewritten. Currently, we adopt both the number of words and the linguistic patterns to identify long sentences. This combined algorithm provides the precision ratio of 52% and the recall ratio of 96% for closed data. The two ratios improved by 9% and 6% respectively, compared with the case when the number of words in a sentence is the only determining factor.

3.2 Generation of Rewriting Candidates

The rewriting rules also generate candidates for rewriting expressions. There are four methods of sentence division:

- 1) Simple division (60%): A sentence is divided into two at the division point inflecting the ending of the first sentence and inserting an appropriate connective at the beginning of the second sentence, where necessary.
- 2) Supplementation of case fillers (27%): A case, such as the subjective case, is supplemented after sentence division.
- 3) Supplementation of verbs (7%): Verbs are supplemented after sentence division.
- 4) Others (6%):

Of the above, method 1) has been implemented and method 2) is under development using semantic trees provided by the information extractor.

4. Experiment and Evaluation

The rewriting tool has been evaluated from two aspects: the knowledge which forms the basis of our pre-editing rules and the tool as a whole, including the interface.

4.1 Evaluation of Rewriting Rules

We have carried out an experiment on the evaluation of our rewriting knowledge using a new test text(211 sentences). The experiment showed almost the same result(62% for the precision ratio and 90% for the recall ratio) as is in Section 3.1.

To evaluate the generation of rewriting candidates

in terms of rewriting positions and rewriting expressions, the precision ratio and the recall ratio of these two have been calculated for the target text, focusing on the sentences for which the tool produced rewriting candidates (i.e., type 1) in Section 3.2.

The precision ratio and the recall ratio for identifying division points are 63% and 100% respectively; the ratios for the rewriting candidates are 10.5% and 93% respectively. The recall ratios for both division points and rewriting expressions exceed 90%, which means that the probability of obtaining the correct positions and candidates are sufficiently high. The precision ratio for substitution generation is low, but 5 to 6 candidates per division point would not be much of a burden on the user.

4.2 Evaluation of the Operating Cost

A preliminary experiment was made on five subjects to evaluate an overall appraisal, including the interface, the rate of reduction in operation time and improvement in the quality of rewriting.

The rate of reduction in rewriting time for the five subjects averages 23%. The recall ratio went up from 83% to 96%, the precision ratio from 96% to 99%. For those subjects whose recall ratio scored high when the tool was not used, the rate of reduction in time also tends to be high. For others whose recall ratio is low when the tool was not used, the rate of reduction in time does not change much, but the recall ratio improved by far. That is, users capable of rewriting without outside help can further shorten the total time using the tool. Moreover, those with low rewriting skills can benefit from the tool to improve rewriting quality.

5. Conclusion

We have proposed an interactive rewriting tool featuring the rewriting candidate generation capabilities to support people in creating machine acceptable sentences. This tool has been applied to MT pre-editing in a Japanese-to-English MT system and showed promising results in the experimental evaluations.

References

Hirakawa, H.; Nogami, H.; and Amano, S. (1991). "EJ/JE Machine Translation System ASTRANSAC--Extensions toward Personalization." *Proc. of MT* SUMMIT III.

Hutchins, W. J. and Somers, H. L. (1992). An Introduction to Machine Translation. ACADEMIC PRESS, pp. 154-155, pp. 324-325.

Kim, Y. B.; Ehara, T.; and Aizawa, T. (1992). "Breaking Long Japanese Sentences Based on Morphological Information."(in Japanese) *Proc. of IPSJ 44th*, vol. 3.

Malone, T. W.; Grant, K. R.; Lai, K.; Rao, R.; and Rosenblitt, D. (1986). "Semi-Structured Messages are Surprisingly Useful for Computer-Supported Coordination." Proc. of Computer-Supported Cooperative Work.