

# Evaluation of Number-Kanji Translation Method of Non-Segmented Japanese Sentences Using Inductive Learning with Degenerated Input Extended Summary

Masafumi Matsuhara<sup>1</sup>, Kenji Araki<sup>2</sup>, Yoshio Momouchi<sup>1</sup>, and Koji Tochinai<sup>2</sup>

<sup>1</sup> Graduate School of Engineering, Hokkai-Gakuen University,  
Minami 26 Nishi 11, Chuo-ku, Sapporo-shi, 064-0926 Japan.

matsu@ai.eli.hokkai-s-u.ac.jp

momouchi@eli.hokkai-s-u.ac.jp

<sup>2</sup> Graduate School of Engineering, Hokkaido University,  
Kita 13 Nishi 8, Kita-ku, Sapporo-shi, 060-8628 Japan.

{araki,tochinai}@media.eng.hokudai.ac.jp

**Abstract.** Our proposed method enables us to promptly and easily input Japanese sentences into a small device. All the keys for input are only 12 keys, which are 0, 1, ..., 9, \* and #. Therefore, we are able to input one *Kana* character per one keystroke. Furthermore, the system based on our method automatically generates the dictionary adapted to the target field because the system automatically acquires words by using inductive learning. The system is improved by its own learning ability.

## 1 Outline

The procedure for our proposed method consists of translation process, proofread process, learning process and feedback process in this order.

A user inputs the string of numbers corresponding to the pronunciation of the intended Japanese sentence by only 12 keys[1][2]. In translation process, the input sentence is translated into *Kanji-Kana* mixed sentence by using the words dictionary. The words in the words dictionary are applied in order of the higher certainty degree[3][4]. The certainty degree is based on the situation of the acquisition of the word, the rate of the correct translation and the appearance degree of neighboring characters[5]. If the translation result has errors, proofread process is performed. The user judges whether it is correct or not and proofreads it. In learning process, words are extracted by comparing the input sentence with its proofread translation result[3][4]. They are compared using their common segments. The extracted words are registered into the words dictionary. In feedback process, the certainty degree for the word in the words dictionary is updated. Thus, this system repeats these processes and improves.

## 2 Evaluation Experiment

The system based on our proposed method has been developed for the experiment. The input data for the experiment are some sections of UNIX MANUAL. The number of characters is 122,000 in the input data. The initial dictionaries are empty for evaluation of adaptability of this method. The system translates each input sentence systematically. The results are evaluated by the translation rates to the number of input characters. The correct rate, erroneous rate and unfixed rate are calculated by each proportion to the number of input characters.

## 3 Results and Considerations

The rate of the correct translation increases as the input data increase. It shows that the system has acquired the words and improved. When the section of input data changes, the rate of correct translation decreases. The reason is that the number of the unregistered words in the words dictionary increases because the field of input data changes. However, the correct rate increases again because this system acquires words for the new fields. Thus, this system adapted to the new field immediately. The final rate of the correct translation is about 85[%].

## Acknowledgements

This work is also partially supported by the Grants from the Government subsidy for aiding scientific researches (No. 09878070, No. 10680367) of Japan Society for the Promotion of Science and the High-Tech Research Center of Hokkai-Gakuen University.

## References

1. Kushler, C.: AAC USING A REDUCED KEYBOARD. *Proc. CSUN98*, Los Angeles, America, Mar. 1998.
2. Higashida, M.: A Fully Automated Directory Assistance Service that Accommodates Degenerated Keyword Input Via Telephone. *Proc. PTC'97*, pp. 167-174, Honolulu, Hawaii, Jan. 1997.
3. Araki, K., Momouchi, Y. and Tochinai, K.: Evaluation for adaptability of Kana-Kanji translation of non-segmented Japanese Kana sentences using inductive learning. *Conference Working Papers of PACLING-II*, pp. 1-7, Brisbane, Australia, Aug. 1995.
4. Araki, K., Takahashi, Y., Momouchi, Y. and Tochinai, K.: Non-Segmented Kana-Kanji Translation Using Inductive Learning. *The Transactions of The Institute of Electronics, Information and Communication Engineers*, Vol. J79-D-II, No. 3, pp. 391-402, Mar. 1996.
5. Nagao, M., Mori, S.: A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese. *proc. COLING94*, pp. 2093-2100, Kyoto, Japan, Aug. 1994.