A Context Sensitive Variant Dictionary for Supporting Variant Selection

Aya Nishikawa, Ryo Nishimura, Yasuhiko Watanabe, Yoshihiro Okada

Ryukoku University, Dep. of Media Informatics, Seta, Otsu, Shiga, Japan
t060606@mail.ryukoku.ac.jp, r_nishimura@afc.ryukoku.ac.jp, watanabe@rins.ryukoku.ac.jp, okada@rins.ryukoku.ac.jp

Abstract

In Japanese, there are a large number of notational variants of words. This is because Japanese words are written in three kinds of characters: kanji (Chinese) characters, hiragara letters, and katakana letters. Japanese students study basic rules of Japanese writing in school for many years. However, it is difficult to learn which variant is suitable for a certain context in official, business, and technical documents because the rules have many exceptions. From the viewpoint of information retrieval, a considerable number of studies have been made on notational variants, however, previous Japanese writing support systems were not concerned with them sufficiently. This is because their main purposes were misspelling detection. Students often use variants which are not misspelling but unsuitable for the contexts in official, business, and technical documents. To solve this problem, we developed a context sensitive variant dictionary. A writing support system based on the context sensitive variant dictionary detects unsuitable variants for the contexts in students' reports and shows suitable ones to the students. This dictionary is based on the idea that context suitable variants are used dominantly in the contexts of official, business, and technical documents. In this study, we first show how to develop a context sensitive variant dictionary by which our system determines which variant is suitable for a context in official, business, and technical documents. Finally, we conducted a control experiment and show the effectiveness of our dictionary.

1. Introduction

In English, there are few words which are spelled in several different ways, such as, color and colour. In contrast, in Japanese, there are a large number of notational variants of words. This is because Japanese words are written in three kinds of characters:

- kanji (Chinese) characters,
- hiragara letters, and
- katakana letters.

Basic rules of Japanese writing are announced by the Cabinet, and Japanese students study them in school for many years. However, it is difficult to learn the rules because they have many exceptions. Take *hikiageru* [pull up] for example. Figure 1 shows the frequencies of notational variants of *hikiageru* [pull up] in the Mainichi newspaper articles. As shown in Figure 1, 引き上げる is the dominant variant of *hikiageru*. In this study, we will use the term *dominant variant* of a word to refer to the most frequent variant of the word. However, as shown in Figure 2, a nondominant variant of *hikiageru*, 引き揚げる, is used dominantly when *hikiageru* is used with *toushi* [investment]. This kind of exceptions often confuse learners of Japanese, not only foreign students but Japanese students. As a result, it is important for students to learn which variant is suitable for a certain context in official, business, and technical documents. To solve this problem, we developed a context sensitive variant dictionary for supporting variant selection. This dictionary is based on the assumption that suitable variants for certain contexts are used dominantly in the contexts of official, business, and technical documents. If the assumption is proper, unsuitable variants for certain contexts can be detected by confirming whether they are used dominantly in the contexts of official, business, and technical documents.

	hiragana	kanji+(1)	kanji+(2)	kanji+(3)
hikiageru	ひきあげる	引きあげる	引き上げる	引き揚げる
[pull up]	1	4	774	146

Figure 1: The frequencies of notational variants of verb "*hikiageru* [pull up]" in the newspaper articles [Mainichi Newspaper (January 2006 – June 2006)]

	hiragana	kanji+(1)	kanji+(2)	kanji+(3)
hikiageru	ひきあげる	引きあげる	引き上げる	引き揚げる
[pull up]	0	0	2	15

Figure 2: The frequencies of notational variants of verb "*hikiageru* [pull up]" in the newspaper articles [Mainichi Newspaper (2005 – 2007)] in the case that the word is used with "*toushi* [investment]"

From the viewpoint of information retrieval, a considerable number of studies have been made on notational variants (Kubomura and Kameda, 2003) (Kouda, 2006) (Bamba et al., 2008), however, spell checkers in Japanese word processor, such as Microsoft word 2007, and previous Japanese writing support systems were not concerned with notational variants sufficiently (Shimomura et al., 1992) (Araki et al., 1993) (Murata and Isahara, 2001). This is because their main purposes were misspelling detection. However, students often use variants which are not misspelling but unsuitable for the contexts in official, business, and technical documents. To solve this problem, we developed a writing support system which detects nondominant variants in students' reports and shows dominant ones to the students (Nishikawa et al., 2009). However, this system is based on a context free variant dictionary. As a result, it is possible that it shows variants which are dominant but un-

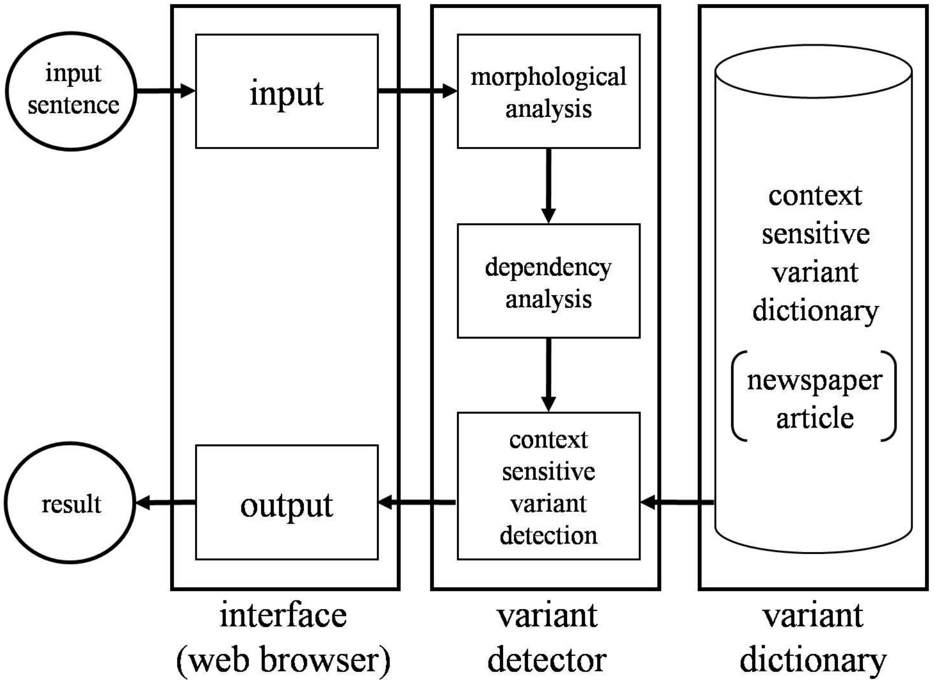

Figure 3: System overview

suitable for the contexts in official, business, and technical documents.

In this study, we first show how to develop a context sensitive variant dictionary by which our system determines which variant is suitable for a certain context in official, business, and technical documents. Finally, we conducted a control experiment and show the effectiveness of our dictionary.

2. Writing support system based on a context sensitive variant dictionary

2.1. System overview

Figure 3 shows the overview of our writing support system based on a context sensitive variant dictionary. As shown in Figure 3, users can access and send input sentences to the system via web browsers by using CGI based HTML forms. Input sentences are segmented into words by using a Japanese morphological analyzer, JUMAN (Kurohashi and Kawahara, 2005a). Then, the dependency relations between the words were analyzed by using a Japanese parser, KNP(Kurohashi and Kawahara, 2005b). Finally, by using the context sensitive variant dictionary, the system confirms whether variants are suitable for the contexts in official, business, and technical documents. When the system detects a unsuitable variant for the context in an input sentence, the system underlines and turns it red, shows the frequency information of the variant in the context, and gives users chances to consider the reasons why they used the variant. In this way, the key to detecting unsuitable variants for the contexts is a context sensitive variant dictionary. In section 2.2., we show how to develop a context sensitive variant dictionary.

2.2. Context sensitive variant dictionary

In order to develop a context sensitive variant dictionary, we expand a context free variant dictionary by adding information of context suitable variants and the contexts. The context free variant dictionary (Nishikawa et al., 2009), which we used and expanded in this study, contains dominant variants of 20929 words which were extracted from 296364 articles published in the Mainichi Newspaper from January 2006 to June 2006 (Mainichi-Shinbun, 2006–2008) credibly by using binomial tests. We expanded this variant dictionary by adding the following kinds of information

- context suitable variants in various contexts
- frequency information of variants in the contexts.

The information was extracted in the next way.
Suppose that word A has a variant which is used in the context that word A is used with word B. We extracted

- the context suitable variant of word A in the context that word A is used with word B, and
- frequency information of variants of word A in the context

in the next steps.

step 1 apply Japanese morphological analysis and dependency analysis to newspaper articles. In this study, we used a Japanese morphological analyzer, JUMAN (Kurohashi and Kawahara, 2005a) and a Japanese parser, KNP(Kurohashi and Kawahara, 2005b).

step 2 From the results of the analyses, extract and count variants of word A which have the dependency relation to word B. In the morphological analysis, JUMAN gives variant labels to variants. Variants of a

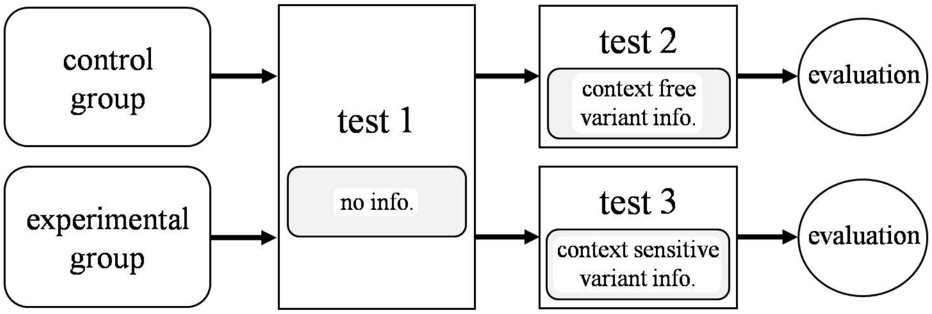

Figure 4: The outline of the experiment

certain word can be detected because JUMAN gives the same variant label to them.

step 3 determine which variant of word A is used dominantly in the context that word A is used with word B. Then, in order to confirm that the variant is a credible context suitable variant, measure the credibility of the context suitable variant by using binomial tests: the variant is regarded as a credible context suitable variant, when the lower limits of one-sided 95% binomial confidence interval of the utilization rates of the variant in the context is more than 0.5.

dominant degree shows how much the dominant variant of a word is used dominantly. Suppose that a word has variant i $(\in I)$ and the utilization rate of variant i is calculated as follows:

$$u_i = \frac{f_i}{\sum_{i \in I} f_i}$$

where u_i and f_i is the utilization rate and frequency of variant i, respectively. The dominant degree of the word is calculated as follows:

$$d = \max_{i \in I} u_i$$

where d is the dominant degree of the word.

In this study, we found 4160735 variants of 25643 words in 1786752 articles published in the Mainichi Newspaper from 2005 to 2007 (Mainichi-Shinbun, 2006–2008). From them, we extracted

- context suitable variants of 13156 words in 40863 contexts
- frequency information of variants in the contexts

by using binomial tests.

3. Experimental results

To evaluate our method, we conducted a control experiment. Figure 4 shows the outline of the experiment. 20 subjects, university students in computer science, were classified into two groups: control group and experimental group. As shown in Figure 4, we conducted test 1 and 2 to the control group, and test 1 and 3 to the experimental group. In these three tests, we gave the same five problems of variant selection with the following kinds of information:

test 1 no information

test 2 context free variant information (Nishikawa et al., 2009)

test 3 context sensitive variant information

Each problem consisted of two sentences, one word of which was underlined, and variant choices of the word. From the variant choices of the underlined word, the subjects were requested to choose one variant which seemed to be suitable for the context in official, business, and technical documents. One sentence in each problem had a context for which the dominant variant was suitable. The other had a context for which the dominant variant was not suitable. For example, the following two sentences were used in a problem of the experiment.

problem 1(a) *kakugi de zeikin wo hikiageru koto ga kettei sareta* [the plan to raise taxes was approved by the Cabinet]

problem 1(b) *New York no sijyo kara toushi wo hikiageru koto ni shita* [we decided to withdraw our investments from the New York market]

The dominant variant of *hikiageru* [pull up] is suitable for the context of problem 1(a), on the other hand, unsuitable for the context of problem 1(b) because *hikiageru* was used with *toushi* [investment]. When subjects in the control group tried to solve problem 1(a) and 1(b) in test 2, they received the frequency information which is shown in Figure 1 and unsuitable for the context of problem 1(b). On the other hand, subjects in the experimental group received context sensitive frequency information. For example, when subjects in the experimental group tried to solve problem 1(b) in test 3, they received the context sensitive frequency information which is shown in Figure 2 and suitable for the context of problem 1(b).

Table 1 shows the choosing rate of variants suitable for the contexts in test 1, 2, and 3. Table 1 shows that the variant selection is a serious problem. In test 1, some subjects chose unsuitable variants for no particular reason and they were totally unaware of doing it. However, Table 1 also implies that students do not have confidence in their variant selection and flexibly change their decisions when the reasons are given to them. Actually, in test 3, five subjects in the experimental group changed their decisions, and two

Table 1: The choosing rate of variants suitable for the contexts

group	test 1	test 2 / 3
control	68%	77%
experimental	73%	81%

other subjects did not change but felt sure of their decisions. Some of them said that they could obey system's advices more simply than teacher's instructions without concrete evidences. On the other hand, in test 2, five subjects in the control group changed their decisions, and two of them selected unsuitable variants for the contexts because of the context free variant information.

4. References

Araki, Ikehara, and Tukahara. 1993. A method for detecting and correcting of characters wrongly substituted, deleted or inserted in japanese strings using 2nd-order markov model. *IPSJ SIG NL*, 93(79):29–35.

Bamba, Shinzato, and Kurohashi. 2008. Development of a large-scale web page clustering system using an open search engine infrastructure tsubaki. *IPSJ SIG NL*, 2008(4):67–74.

Kouda. 2006. Search method of variant notations on a science and technology document retrieval system. *IPSJ SIG NL*, 2006(118):5–10.

Kubomura and Kameda. 2003. Information retrieval system with abilities of processing katakana-allographs. *Transactions of IEICE*, J86-D-II(3):418–428.

Kurohashi and Kawahara. 2005a. Juman manual version 5.1. *Kyoto University*.

Kurohashi and Kawahara. 2005b. Knp manual version 2.0. *Kyoto University*.

Mainichi-Shinbun. 2006–2008. Mainichi shinbun cd-rom data set 2005, 2006, and 2007. *Nichigai Associates Co.*

Murata and Isahara. 2001. Extraction of negative examples based on positive examples: automatic detection of misspelled japanese expressions and relative clauses that do not have case relations with their heads. *IPSJ SIG NL*, 2001(69):105–112.

Nishikawa, Nishimura, Watanabe, and Okada. 2009. Writing support system dealing with notational variant selection. *CSEDU 2009*, pages 73–80.

Shimomura, Namiki, Nakagawa, and Takahashi. 1992. A method for detecting errors in japanese sentences based on morphological analysis using minimal cost path search. *Transanctions of IPSJ*, 33(4):457–464.

Acknowledgement This research has been supported partly by the Grant-in-Aid for Scientific Research (C) under Grant No.20500106.