Generating "A for Alpha" When There Are Thousands of Characters

Hiroaki KAWASAKI[1] *Ryohei SASANO*[2]
Hiroya TAKAMURA[2] *Manabu OKUMURA*[2]
(1) Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology
(2) Precision and Intelligence Laboratory, Tokyo Institute of Technology
kawa@lr.pi.titech.ac.jp, {sasano,takamura,oku}@pi.titech.ac.jp

ABSTRACT
The phonetic alphabet enables people to dictate letters of the alphabet accurately by using representative words, i.e., A for Alpha. Japanese kanji (idiographic Chinese characters) vastly outnumber the letters of the Roman alphabet, and thus Japanese requires an explanatory reading like a phonetic alphabet. We call the explanatory reading of a kanji a "distinctive explanation." Most kanji characters have their homophones, and the role of the distinctive explanations is to enable users to identify a specific kanji character only by listening to the explanation. In this paper, we propose a corpus-based method for automatically generating distinctive explanations for a kanji, in which information about familiarity and homophones of kanji are taken into consideration. Through the kanji-identification experiments, we show that the quality of the explanations generated by the proposed method is higher than that of the manually crafted distinctive explanations.

KEYWORDS: phonetic alphabet, distinctive explanation for a kanji.

Proceedings of COLING 2012: Technical Papers, pages 1411–1424,
COLING 2012, Mumbai, December 2012.

1 Introduction

Japanese has three types of characters: hiragana, katakana, and kanji (Chinese characters). While hiragana and katakana characters are phonograms, kanjis are ideograms, each of which usually has several readings. Most kanjis have homophones, and thus it is difficult to identify a kanji only by its reading. However, sometimes we need to identify a kanji verbally. For example, screen readers need to enable the users, especially visually impaired people who have difficulty seeing things, to identify a kanji only by sound. When we talk over the telephone, we have to exchange information, such as proper names, only by our voice. In such cases, we explain a kanji by using not only its reading but also its properties, compositions, and so on, to reduce the ambiguity. In this paper, we call such an explanation a "distinctive explanation for a kanji" and propose a method for generating distinctive explanations automatically.

One concept similar to a distinctive explanation is the phonetic alphabet, e.g., A for Alpha, B for Bravo, and C for Charlie. While a distinctive explanation explains kanjis, the phonetic alphabet explains letters and numbers. One of the major differences between them is the number of target characters. Whereas the phonetic alphabet for English deals with only 26 letters and 10 numbers, the distinctive explanation for kanjis deals with thousands of kanji characters.

The distinctive explanation leverages various aspects of the target kanji such as its Chinese reading, its Japanese reading, and its radicals, to reduce ambiguity. Each kanji has several readings, some of which are derived from Chinese readings and the others are Japanese traditional readings. If a kanji has a distinctive reading, the reading can be used to generate the distinctive explanation. Some kanjis can be divided into left and right parts and we can explain a kanji without ambiguity by using them. For example, the kanji "評 (*hyou*)" can be divided into "言 (*gon*)" and "平 (*hei*)." By listening to the information of "言 (*gon*)" and "平 (*hei*)," we can identify "評 (*hyou*)." Words including the target kanji, such as "購入 (*kou-nyū*, purchase)" for "購 (*kou*)," are also effective in identifying the target kanji because they can reduce the ambiguity.

Some screen readers (Lazar et al., 2007) already have functions for outputting distinctive explanations for kanjis. A screen reader is a software application for visually impaired people that reads aloud a text on a computer screen. This function enables visually impaired people to use e-mail, read news, view Web pages, and operate other complex applications. However, the existing screen readers use some distinctive explanations that do not make a target kanji easily identifiable, such as "*aya* for[1] *aya-ori* (綾織, twill)" for "綾 (*aya*)." "綾織 (*aya-ori*, twill)" is not a word with which most people are familiar, and thus we cannot identify the target kanji "綾 (*aya*)" easily. Watanabe et al. (2003) pointed out that the main factors that prevent the users from identifying the target kanji are the low familiarity and the homophones of the words used in the distinctive explanations.

Vocabulary and word familiarity vary among age groups, regions, and social backgrounds. It is hard work to remake a distinctive explanation for each kanji in accordance with changes in the target audience. We try to automate both the acquisition of vocabulary and word familiarity, and the generation of the distinctive explanations.

[1]In fact, distinctive explanations are expressed by using a Japanese word "*no*" as in "*kou-nyū no kou*." "*no*" is a Japanese postposition that can represent a wide range of semantic relations. It is similar to "for" in English. In this paper, we therefore refer to distinctive explanations by using "for" as in "*kou* for *kou-nyū*."

In this paper, we propose a method for automatically generating distinctive explanations for a kanji, and aim to improve the kanji identification rate. Our system automatically generates distinctive explanations using the knowledge of familiarity and homophones derived from a large text corpus. Automatic methods for generating the distinctive explanation can easily adapt to the users.

2 Distinctive explanation for kanji

With the growth of computers, screen readers that have functions for producing voice outputs of distinctive explanations have become popular among visually impaired people. Accordingly, people argued over the problem of what distinctive explanations should be outputted from screen readers.

Ooyama et al. (1996) proposed a spoken explanation generator, PLANET. This system can explain a kanji, especially those used in peoples' names, only by sound. When explaining, the system uses the information such as words containing the target kanji and the components of the kanji.

Watanabe et al. (2005a) manually produced distinctive explanations on the basis of children's vocabulary familiarity obtained by a kanji dictation survey. When producing distinctive explanations, they prioritized words that had higher familiarity than others and no homophones. Moreover, they avoided using negative expressions and English words like "*kin*[2] *for* gold," so that generated distinctive explanations would be suitable for elementary school students. The identification rate for their generated distinctive explanations was 14.1% higher than that for the existing distinctive explanations in the experiments of kanji dictation involving elementary school students.

Nishida et al. (2005) proposed distinctive explanations based on the meanings of a kanji. For example, the traditional distinctive explanation of "情報 (*jou-hou*, information)" was "*jou* for *jou-netsu* (情熱, passion)" and "*hou* for *hou-koku* (報告, report)." On the other hand, in distinctive explanations based on the meanings, "情報 (*jou-hou*, information)" was explained as "*i-n-fo-mē-sho-n* (インフォメーション, information)," "*chou-hou* (諜報, intelligence)," or "*hi-mitsu-jou-hou* (秘密情報, confidential information)." Experimental results showed no differences between the identification rates of the traditional distinctive explanations and explanations based on the semantics. Since those main target words are those that appear in a thesaurus, we cannot easily compare them and our distinctive explanations.

Watanabe et al. (2005b) classified in detail the composition of the traditional distinctive explanations derived from screen readers. Distinctive explanations can be classified into three types in accordance with their configuration:

Type 1 consists of a word that includes the target kanji and the reading of the target kanji in the word.

"*kou* for *kou-nyū* (購入, purchase)" for "購 (*kou*)"

Type 2 consists of the distinctive reading of the target kanji. This type uses the reading that other kanjis never have.

"*sakura* (桜, cherry blossom)" for "桜 (*sakura*)"

[2]In Japanese, *kin* means gold.

Type 3 uses the components of radicals of the target kanji or consists of meanings of the target kanji that forms a word only by itself.

"*kawa* with *sanzui*[3]" for "河 (*kawa*)"

They reported that Type 1 distinctive explanations were most common. Type 1 is suitable for use in statistical treatment, and therefore we aim to generate Type 1 distinctive explanations automatically in this paper.

Watanabe et al. also investigated the factors that make it difficult to identify the target kanji (Watanabe et al., 2003). We enumerate the major factors reported in their work:

Factor 1 The low familiarity of words such as "千代紙 (*chi-yo-gami*, Japanese paper with colored figures)."

Factor 2 The presence of homophones such as "購買 (*kou-bai*, purchase)" and "勾配 (*kou-bai*, gradient)."

Factor 3 The target kanji itself is difficult, such as "爾 (*ji*, thou)."

Factors 1 and 2 can be improved by selecting a suitable word for the distinctive explanation, but Factor 3 cannot because of the difficulty of the target kanji itself. Our study aims to improve the rate of identifying the target kanji from the distinctive explanation, and hence we focus on Factors 1 and 2.

3 Automatic generation of distinctive explanation for kanji

We propose an interactive system that automatically generates distinctive explanations for a kanji. Figure 1 shows the overview of our system. The first step outputs one Type 1 distinctive explanation. If the user cannot recall a kanji, the second step outputs another distinctive explanation.

The first step uses a single word that has high familiarity and few homophones. However, some kanjis are hard to identify from one word unambiguously; for example, "科 (*ka*)." While the most common words that include "科 (*ka*)" are "科学 (*ka-gaku*, science)," "教科 (*kyou-ka*, subject)," and "単科 (*tan-ka*, single subject)," all have several homophones such as "化学 (*ka-gaku*, chemistry)" for "科学 (*ka-gaku*)," "強化 (*kyou-ka*, reinforcement)" for "教科 (*kyou-ka*)," and "炭化 (*tan-ka*, carbonization)" and "単価 (*tan-ka*, unit price)" for "単科 (*tan-ka*, single subject)," and thus it is hard to identify "科 (*ka*)" unambiguously from a Type 1 distinctive explanation. For such a kanji, our system proceeds to the second step, and generates another distinctive explanation for the kanji by using a word that ensures a high kanji identification rate when combined with the word presented by the first step of our system.

The first example of the system's input in Figure 1 is the kanji "購 (*kou*)." The system outputs the distinctive explanation "*kou* for *kou-nyū*" by our first step and describes it to the user. The user identifies the correct kanji "購 (*kou*)" from the distinctive explanation. The second example of the system's input is the kanji "科 (*ka*)." The system outputs the distinctive explanation "*ka* for *ka-gaku*" by our first step and describes it to the user. The user cannot identify

[3] *sanzui* means "氵."

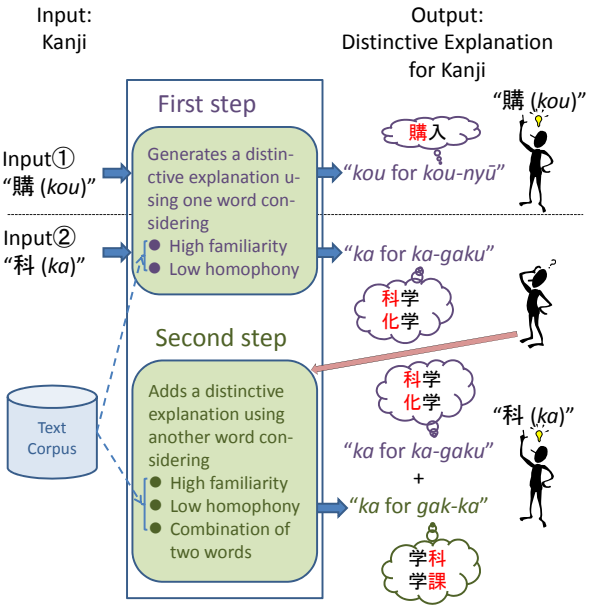

Figure 1: The overview of our system. If “購 (*kou*)” is input, the first step outputs “*kou* for *kou-nyū*.” In this case, the user will think of the correct kanji since there is no ambiguity. If “科 (*ka*)” is input, the first step outputs “*ka* for *ka-gaku*.” However, there are plural candidates such as “科 (*ka*)” and “化 (*ka*).” In such case, the user asks our system to generate an additional distinctive explanation. The second step then outputs "*ka* for *gak-ka*," and the user can identify the correct kanji.

the correct kanji because the reading “*ka-gaku*” has many homophones. Then the user asks our system to generate an additional distinctive explanation. Our system outputs the second distinctive explanation “*ka* for *gak-ka*,” and the user identifies the correct kanji “科 (*ka*)” from the two distinctive explanations.

3.1 Generation of distinctive explanations in the first step

The first step generates a Type 1 distinctive explanation as follows:

i. Extract words from the corpus that have more than two characters and include the target kanji.

ii. Calculate the score for each word w:

$$score_1(w) \triangleq p(w)^{\alpha} \cdot u_1(w)^{\beta}, \tag{1}$$

where

$p(w)$: the probability of w in the corpus,

$u_1(w)$: the ratio of the frequency of w to the frequency of all words in the corpus that have the same reading as w,

α: a parameter that reflects the importance of familiarity ($\in [0, 1]$),

β: a parameter that reflects the importance of not having homophones ($\in [0, 1]$).

The probability $p(w)$ is calculated as follows:

$$p(w) \triangleq \frac{c(w)}{\sum_{w' \in W} c(w')}, \tag{2}$$

where

$c(w)$: the frequency of w in the corpus,

W: the set of words that appear in the corpus.

The ratio $u_1(w)$ is calculated in the following way:

$$u_1(w) \triangleq \frac{c(w)}{\sum_{w' \in H(w)} c(w')}, \tag{3}$$

where

$H(w)$: the set of words that have the same reading as w.

iii. Select the word with the highest score and then use it to generate a Type 1 distinctive explanation.

In Equation (1), $p(w)$ represents the degree of familiarity and $u_1(w)$ represents the degree of uniqueness. For example, the most common words that include "購 (*kou*)" are "購読 (*kou-doku*, subscription)," "購入 (*kou-nyū*, purchase)," and "購買 (*kou-bai*, purchase)." While "購読 (*kou-doku*, subscription)" has higher probability than the other words, it has homophones such as "鉱毒 (mining pollution)." $u_1(w)$ functions to reduce the score of such ambiguous words. As a result, our system prioritizes the output of the distinctive explanation using the word "購入 (*kou-nyū*, purchase)" rather than "購読 (*kou-doku*, subscription)."

3.2 Generation of distinctive explanations in the second step

The second step generates an additional distinctive explanation that reduces ambiguity when combined with the distinctive explanation by the first step.

i. Extract words from the corpus that have more than two characters and include the target kanji.

ii. Give a score for each pair of the word selected in the first step w_1, and a word extracted in i. The score is calculated in the following way:

$$score_2(w_1, w) \triangleq score_1(w_1) \cdot score_1(w) \cdot u_2(w_1, w)^{\gamma}, \tag{4}$$

where

$u_2(w_1, w)$: the ratio of the number of the target kanjis to the number of the kanjis that we can recall from the pairs composed of w_1 and w,

γ: a parameter that reflects the importance of combination ambiguity ($\in [0, 1]$).

The ratio $u_2(w_1, w)$ is calculated in the following way:

$$u_2(w_1, w) \triangleq \frac{\min(c(w_1), c(w))}{\sum_{(w_1', w') \in C} \min(c(w_1'), c(w'))}, \tag{5}$$

where

C: the set of candidate pairs of words that have the same reading as w_1 and w and can make the user recall a kanji.

iii. Select the word with the highest score and then generate a distinctive explanation by using w besides w_1.

Equation (4) consists of a product of $score_1(w_1)$, $score_1(w)$, and $u_2(w_1, w)^{\gamma}$. $u_2(w_1, w)^{\gamma}$ represents the uniqueness when using two words. For example, the distinctive explanation "*ka* for *ka-gaku*" and "*ka* for *tan-ka*" evoke at least two kanjis: "科 (*ka*)" and "化 (*ka*)." The candidate kanjis for the distinctive explanation "*ka* for *ka-gaku*" are "科 (*ka*)" from "科学 (*ka-gaku*, science)" and "化 (*ka*)" from "化学 (*ka-gaku*, chemistry)." Similarly, those for "*ka* for *tan-ka*" are "科 (*ka*)" from "単科 (*tan-ka*, single subject)" and "化 (*ka*)" from "炭化 (*tan-ka*, carbonization)" and "価 (*ka*)" from "単価 (*tan-ka*, unit price)." Thus we have $C = \{($ 科学, 単科 $), ($ 化学, 炭化 $)\}$. The term $u_2(x, y)^{\gamma}$ reduces the scores of such ambiguous distinctive explanations. Our system outputs distinctive explanations that are less ambiguous, such as "*ka* for *ka-gaku*" and "*ka* for *gak-ka*."

4 Experiments

4.1 Experimental Settings

We used three corpora in experiments:

- Google Japanese N-gram corpus (Google corpus)
- Yomiuri newspaper corpus (Yomiuri corpus)
- Balanced Corpus of Contemporary Written Japanese (BCCWJ)

The Google Japanese N-gram corpus (Kudo and Kazawa, 2007) was constructed from 20 billion Japanese sentences on the Web and consists of 255 billion words. The Yomiuri newspaper corpus consists of 400 million words in Yomiuri newspaper articles from 1991 to 2004. The

BCCWJ (Maekawa, 2008) is a balanced corpus of 100 million words of contemporary written Japanese.

We used MeCab (Kudo et al., 2004), an open-source morphological analyzer, to separate the corpus into words. We applied the IPA dictionary with the parameter estimated by Conditional Random Fields (CRF) based on IPA corpus[4]. We also used the Simple Kana to Kanji conversion program (SKK) dictionary[5] to utilize words longer than those in the IPA dictionary. This is because the words in the IPA dictionary tend to be too short for our method. For example, although we want to use "炒め物 (*ita-me-mono*, fried food)" to generate a distinctive explanation for "炒 (*ita*)," there is no entry for "炒め物" in the IPA dictionary, and thus "炒め物" is divided into two words: "炒め (*ita-me*, fried)" and "物 (*mono*, object)." On the other hand, there is an entry for "炒め物" in the SKK dictionary, and in such cases we consider the word as a candidate word for generating distinctive explanations.

We used the kanjis selected from top 2,000 frequently occurring kanjis in the Google corpus for experiments. This is because we want to focus on the performance of generating distinctive explanations, and thus we want to ignore errors caused by Factor 3: the target kanji itself is difficult. Note that since the total frequency of the 2,000 kanjis covered more than 99% kanji occurrences, our experimental setting is very practical. We set the parameters (α, β, γ) to $(0.1, 1.0, 1.0)$ in accordance with the results of a preliminary experiment.

4.2 Comparison of the three corpora

We first conducted a preliminary experiment to evaluate distinctive explanations generated by the first step to examine which corpus is the most preferable for our method. We prepared four distinctive explanations for each kanji: three are generated by using each corpus (Google corpus, Yomiuri corpus, and BCCWJ) and one is the distinctive explanation used in the screen reader PC-Talker XP for comparison. We randomly selected 100 kanjis for an evaluation from the 2,000 most frequently occurring kanjis that had Type 1 distinctive explanations in PC-Talker XP such as "*kou* for *kou-nyū* (購入, purchase)" for "購 (*kou*)."

The distinctive explanations were evaluated by eight human subjects. These distinctive explanations were written on paper, randomly shuffled, and shown to the subjects. Each subject was shown only one distinctive explanation for each kanji. Since there were eight human subjects and four types of distinctive explanations were generated for each kanji, each distinctive explanation was evaluated by two subjects. The subjects were requested to think of the most likely kanji from the presented distinctive explanation and to choose one from the following choices:

a. I thought of a kanji that matched the target kanji.

b. I thought of a kanji that did not match the target kanji.

c. I could not think of any kanji.

We did not conduct a kanji dictation test when we evaluated the distinctive explanations. Japanese speakers cannot always write out kanjis that they can identify, probably because of

[4]http://en.sourceforge.jp/projects/ipadic/
[5]http://openlab.ring.gr.jp/skk/index.html

	Google corpus	Yomiuri corpus	BCCWJ	PC-Talker XP
a	179	170	185	185
b	15	15	9	9
c	6	15	6	6
IR[%]	89.5	85.0	92.5	92.5

Table 1: The identification rates for three corpora.

the spread of computers, which let us input kanjis simply by selecting the correct one instead of actually writing it. This situation is similar to the fact that some English speakers cannot always spell familiar words correctly.

We calculated the identification rate (IR), i.e., the percentage of successfully identified kanji, for each corpus, as follows:

$$\mathrm{IR} = \frac{n(\mathbf{a})}{n(\mathbf{a}) + n(\mathbf{b}) + n(\mathbf{c})} \times 100\ [\%] \tag{6}$$

where $n(x)$ is the number of times choice x is selected.

Table 1 shows the identification rates for three corpora. PC-Talker XP and our system based on BCCWJ achieved the highest identification rate (92.5%). We conducted the McNemar's test and confirmed that there were significant differences at the 0.05 significance level between our system with Yomiuri corpus and PC-Talker XP, and between our system with Yomiuri corpus and with BCCWJ.

Table 2 shows the examples of distinctive explanations and their evaluation. Distinctive explanations generated by the Yomiuri corpus (newspaper articles) tend to use difficult words such as "*gai* for *gai-bou* (外貌, exterior)" generated for "貌 (*bou*)" and "*gi* for *yo-gi-nai* (余儀無い, unavoidable)" generated for "儀 (*gi*)," while there are easier words such as "美貌 (*bi-bou*, beauty)" and "儀式 (*gi-shiki*, ceremony)." This tendency would be one of the factors of the lower identification rate of distinctive explanations generated by the Yomiuri corpus.

4.3 Evaluation of the proposed method

We then evaluated the proposed method with both steps. We used the output of PC-Talker XP as a comparison. On the basis of the results in Sec. 4.2, we used BCCWJ as the corpus in this experiment. We used 100 kanjis randomly selected from the 2,000 most frequent kanjis that were not limited to kanjis that had Type 1 distinctive explanations in PC-Talker XP. Table 3 shows the number of kanjis for each type used in PC-Talker XP.

We evaluated 200 distinctive explanations: 100 generated from our method and 100 extracted from PC-Talker XP. Sixty subjects were each shown 50 distinctive explanations. Thus each distinctive explanation was evaluated by 15 subjects.

When evaluating our system, we first asked the subjects to think of a kanji from the distinctive explanation generated by the first step described in the paper. If the subjects could not think of a kanji, we asked the subjects to look at the distinctive explanation generated by the second step and to think of a kanji. After that, we asked the subjects to choose one from the following:

Kanji	Google corpus	Yomiuri corpus	BCCWJ	PC-Talker XP
儀	"*gi* for *sou-gi*" 葬儀 (2/2) funeral	"*gi* for *yo-gi-na-i*" 余儀無い (0/2) unavoidable	"*gi* for *gi-shiki*" 儀式 (2/2) ceremony	"*gi* for *gi-shiki*" 儀式 (2/2) ceremony
貌	"*bou* for *bi-bou*" 美貌 (2/2) beauty	"*bou* for *gai-bou*" 外貌 (0/2) exterior	"*bou* for *bi-bou*" 美貌 (2/2) beauty	"*bou* for *bi-bou*" 美貌 (2/2) beauty
感	"*kan* for *kan-ji*" 感じ (0/2) feeling	"*kan* for *kan-jiru*" 感じる (2/2) feel	"*kan* for *kan-jiru*" 感じる (2/2) feel	"*kan* for *kan-shin-suru*" 感心する (2/2) be impressed
遥	"*you* for *you-hai*" 遥拝 (0/2) worshipping from afar	"*you* for *you-hai*" 遥拝 (0/2) worshipping from afar	"*you* for *you-hai*" 遥拝 (0/2) worshipping from afar	"*haru* for *haru-ka-kanata*" 遥か彼方 (2/2) far away
餅	"*hei* for *sen-bei*" 煎餅 (1/2) rice cracker	"*mochi* for *kiri-mochi*" 切餅 (1/2) sliced cracker	"*hei* for *sen-bei*" 煎餅 (2/2) rice cracker	"*mochi* for *mochi-tsuki*" 餅つき (1/2) mochi pounding
欄	"*ran* for *kū-ran*" 空欄 (2/2) blank	"*ran* for *ran-kan*" 欄干 (0/2) parapet	"*ran* for *ran-kan*" 欄干 (1/2) parapet	"*ran* for *ran-gai*" 欄外 (2/2) margin
点	"*ten* for *kyo-ten*" 拠点 (1/2) stronghold	"*ten* for *kyo-ten*" 拠点 (1/2) stronghold	"*ten* for *kan-ten*" 観点 (0/2) standpoint	"*ten* for *ten-sū*" 点数 (2/2) score
輸	"*yu* for *yu-nyū*" 輸入 (2/2) importation	"*yu* for *yu-nyū*" 輸入 (2/2) importation	"*yu* for *yu-nyū*" 輸入 (2/2) importation	"*yu* for *yu-shutu-su-ru*" 輸出する (2/2) export

Table 2: Examples of distinctive explanations and their evaluation. "$(n/2)$" means n subjects out of two chose **a**.

- I thought of a specific kanji from only the first distinctive explanation, and the kanji was
  $\mathbf{a_1}$. correct, $\mathbf{b_1}$. wrong.
- I thought of a specific kanji from the first and the second distinctive explanation, and the kanji was
  $\mathbf{a_2}$. correct, $\mathbf{b_2}$. wrong.
- **c**. I could not think of any kanji.

To evaluate our whole system, we regarded $\mathbf{a_1}$ and $\mathbf{a_2}$ as positive answers and calculated the identification rate (IR_2) as follows:

$$\mathrm{IR}_2 = \frac{n(\mathbf{a_1}) + n(\mathbf{a_2})}{n(\mathbf{a_1}) + n(\mathbf{b_1}) + n(\mathbf{a_2}) + n(\mathbf{b_2}) + n(\mathbf{c})} \times 100.\ [\%] \tag{7}$$

We also evaluated the distinctive explanations generated only by the first step in our system for comparison. For this evaluation, we regarded $\mathbf{a_1}$ as a positive answer and calculated the identification rate (IR_1) as follows:

$$\mathrm{IR}_1 = \frac{n(\mathbf{a_1})}{n(\mathbf{a_1}) + n(\mathbf{b_1}) + n(\mathbf{a_2}) + n(\mathbf{b_2}) + n(\mathbf{c})} \times 100.\ [\%] \tag{8}$$

Type	#	Example
Type 1	93	"*ka* for *ka-zei-su-ru* (課税する, tax)" "*ken* for *ken-ka-su-ru* (喧嘩する, fight)"
Type 2	0	-
Type 3	7	"*yorokobi-wo-imi-suruka* (happiness)" "*tsuchi-wo-hutatu-kasaneta kei* (to stack soil on soil)"

Table 3: The number of kanjis for each type used in PC-Talker XP.

	Our system using BCCWJ	PC-Talker XP
$\mathbf{a_1}$	1,181	1,301
$\mathbf{b_1}$	28	58
$\mathbf{a_2}$	163	-
$\mathbf{b_2}$	22	-
c	106	141
IR[%]	IR_1: 78.7, IR_2: 89.6	IR_{SR}: 86.7

Table 4: Evaluation results of our system outputs and distinctive explanations in screen reader.

To evaluate the distinctive explanations in PC-Talker XP, we asked the subjects to choose one from the following options:

a. I thought of a specific kanji that was correct.

b. I thought of a specific kanji that was wrong.

c. I could not think of any kanji.

For evaluating the screen reader, we regarded **a** as a positive answer and calculated the identification rate (IR_{SR}) as follows:

$$IR_{SR} = \frac{n(\mathbf{a})}{n(\mathbf{a}) + n(\mathbf{b}) + n(\mathbf{c})} \times 100. \ [\%] \tag{9}$$

Table 4 shows the results. We confirmed that our whole system outperformed PC-Talker XP. We conducted the McNemar's test and confirmed that our whole system (IR_2) and PC-Talker XP significantly differed at the 0.05 level. Distinctive explanations generated by our system seem to be longer and take more time to listen to than those of PC-Talker XP. However, users do not always need to hear the entire distinctive explanation of our system to think of a kanji. Table 5 shows the average length of distinctive explanations shown to the subjects. In 80.6 %[6] of cases, the target kanjis were correctly thought of in the first step. In addition, the average length of the first step's output was shorter than that of distinctive explanations of PC-Talker XP. The average length of our system output was 8.14, which was shorter than that of PC-Talker XP, and thus the comparison of (IR_2) and (IR_{SR}) is fair.

The identification rate of the first step (IR_1) was lower than both those of the screen reader (IR_{SR}) and the rate in Sec. 4.2 (IR). We think there are two reasons for this. The first

[6] $(1,181 + 28)/1,500 = 0.806$

	First step	Second step	The average length
Our system	6.80	13.93	8.14
PC-Talker XP	-	-	8.96

Table 5: The average lengths of distinctive explanations shown to the subjects.

Kanji	Our system		PC-Talker XP
	First step	Second step	
悟	"*go* for *kaku-go*" 覚悟 (9/15) preparation	"*sato* for *sato-ri*" 悟り (14/15) enlightenment	"*go* for *kaku-go, sato-ru*" 覚悟; 悟る (13/15) preparation; to realize
課	"*ka* for *ka-dai*" 課題 (13/15) assignment	"*ka* for *ka-zei*" 課税 (15/15) taxation	"*ka* for *ka-zei-suru*" 課税する (6/15) tax
灌	"*kan* for *yu-kan*" 湯灌 (1/15) wash a dead body	"*kan* for *kan-gai-you-sui*" 灌漑用水 (7/15) irrigation	"*kan* for *kan-gai-suru, soso-gu*" 灌漑する; 灌ぐ (1/15) irrigate; pour
藍	"*ran* for *ga-ran*" 伽藍 (0/15) temple	"*ai* for *ai-hara*" 藍原 (4/15) family name	"*ai* for *ai-iro*" 藍色 (12/15) indigo blue
圭	"*kei* for *kei-ji-rou*" 圭二郎 (2/15) first name	"*kei* for *kei-ichi*" 圭一 (4/15) first name	"*tsuchi-wo-hutatu-kasaneta kei*" (11/15) to stack soil on soil
嘩	"*ka* for *hū-hu-gen-ka*" 夫婦喧嘩 (4/15) marital quarrel	"*ka* for *ō-gen-ka*" 大喧嘩 (5/15) big fight	"*ken* for *ken-ka-suru*" 喧嘩する (1/15) fight
嘉	"*ka* for *ka-ei*" 嘉永 (0/15) era name	"*ka* for *ka-de-na*" 嘉手納 (0/15) place-name	"*yoroko-bi-wo-i-mi-suru ka*" 嘉 (1/15) that means happiness

Table 6: Examples of distinctive explanations generated by the whole our system and distinctive explanations in PC-Talker XP and their evaluation. "$(n/15)$" means n subjects out of 15 chose a positive answer.

is the differences between evaluation methods. While the prior evaluation (IR) contained the possibility of positive evaluation when subjects thought of multiple kanjis, this evaluation (IR_1) did not. The second is the use of different kanjis. While kanjis are limited to those that have Type 1 distinctive explanations in PC-Talker XP in the prior evaluation (IR), all kanjis were allowed in this evaluation. Even for the kanji unsuitable for the Type 1 distinctive explanation, our system has to output Type 1 distinctive explanations.

Table 6 shows examples of distinctive explanations and their evaluation. We confirmed that our system generates a better distinctive explanation for "課 (*ka*)" and "灌 (*kan*)." In the case of "課 (*ka*)," 13 subjects out of 15 successfully identified "課 (*ka*)" from distinctive explanations generated by the first step. However, two subjects could not identify the target kanji. This would be because our system outputted the distinctive explanation using the word "課題 (*ka-dai*, assignment)," which has homophones such as "過大 (*ka-dai*, excessive)" and "仮題 (*ka-dai*, a tentative title)." Since the remaining two identified the correct kanji by using distinctive explanations generated by the second step, we confirmed the effectiveness of the

proposed two-step method. On the other hand, only six subjects identified kanji from distinctive explanations of the screen reader. The cause of the low identification rate of the screen reader may be that subjects thought of "*ka-sei-suru* (加勢する, assist)" instead of "*ka-zei-suru* (課税する, tax)." In the case of "灌 (*kan*)," only one subject out of 15 identified "灌 (*kan*)" from our first step or the screen reader. However, when other subjects looked at distinctive explanations generated by the second step, seven identified "灌 (*kan*)." It can be inferred from these results that "灌 (*kan*)" is difficult to identify but our two-step approach is effective in such a case.

Conversely, the examples where our system was worse than the screen reader were the cases of "藍 (*ai* or *ran*)" and "圭 (*kei*)." In the case of "藍 (*ai* or *ran*)," our first step could not make anyone identify the kanji. In addition, even our second step made only four subjects identify the kanji. However, the screen reader succeeded in making 12 subjects identify the kanji. This is because our system cannot capture the specific features of "藍 (*ai* or *ran*):" in Japanese, "藍色 (*ai-iro*, indigo blue)" is a color. While the screen reader used words related to the color, our system used "伽藍 (*ga-ran*, temple)" and "藍原 (*ai-hara*, which is a Japanese family name)." The neither word includes information about the color. Such kanji-specific information is important but our system cannot use it well. For "圭 (*kei*)," our system used ambiguous words, and most subjects failed to identify the kanji[7]. In contrast, the screen reader achieved a high identification rate for this kanji, by using the distinctive explanation "Write 土 on top of 土 (*tsu-chi*, soil)." The use of kanji components or radicals as in this example by the screen reader, on top of our method, will further improve the identification performance.

Although we selected the top 2,000 frequent kanjis for the candidates of evaluation in order to eliminate the difficult kanji, some difficult kanjis still appeared. For example, in the case of "嘉[8] (*ka*)," our system and the screen reader made barely any subjects identify the kanji. We think that this is because "嘉 (*ka*)" itself is difficult.

Conclusion

In this paper, we proposed a method for automatically generating distinctive explanations for a kanji using a text corpus. The proposed method took into account familiarity and homophones of kanjis. As a result of human evaluation, we confirmed that distinctive explanations generated by our system outperformed those in existing screen readers.

Our future work involves incorporation of intonation, application to Chinese, and user adaptation. Intonation of words can help generate good distinctive explanations. For example, "橋 (*hashi*, bridge)" and "箸 (*hashi*, chopstick)" have the same readings but different intonations, so we think intonation can be a clue for identifying a kanji. Kanjis are used in not only Japanese but also Chinese. Since our proposed method is language-independent, our method can be applied to distinctive explanations for Chinese. Finally, we are considering generating distinctive explanations that consider the user attributes. For example, users who have studied the law will be familiar with legal terms but not medical terms. To adapt our system to different users, we can select a corpus that is suitable for them.

[7] "*Kei-ji-rou*" has homophones such as "慶二郎," "敬二郎," and "啓二郎" and "*kei-ichi*" have homophones such as "恵一," "慶一," and "啓一." All are Japanese male names.

[8] "嘉 (*ka*)" means happiness, but this kanji is rarely used.

References

Kudo, T. and Kazawa, H. (2007). Japanese web n-gram corpus version 1. *Google/Linguistic Data Consortium*.

Kudo, T., Yamamoto, K., and Matsumoto, Y. (2004). Applying conditional random fields to Japanese morphological analysis. *In Proceedings of EMNLP*, pages 230–237.

Lazar, J., Allen, A., Kleinman, J., and Malarkey, C. (2007). What frustrates screen reader users on the web: A study of 100 blind users. *Int. J. Hum. Comput. Interaction*, pages 247–269.

Maekawa, K. (2008). Balanced corpus of contemporary written Japanese. *Proceedings of the 6th Workshop on Asian Language Resources*, pages 101–102.

Nishida, M., Horiuchi, Y., and Ichikawa, A. (2005). Kanji transformation using semantic information for blind people. *IEICE technical report. Welfare Information technology (In Japanese)*, 105(186):1–6.

Ooyama, Y., Asano, H., and Matsuoka, K. (1996). Spoken-style explanation generator for japanese kanji using a text-to-speech system. 3:1369–1372.

Watanabe, T., Teruyoshi, F., Watanabe, B., Sawada, M., and Kamata, K. (2003). A consideration on shosaiyomi (explanatory expressions) used in screen readers for visually-impaired persons. *Technical report of IEICE. HCS (In Japanese)*, 102(599):25–28.

Watanabe, T., Watanabe, B., Okada, S., Yamaguchi, T., Oosugi, N., and Sawada, M. (2005a). A study on shosaiyomi of screen readers kanji writing test using newly devised shosaiyomi. *IEICE technical report. Welfare Information technology (In Japanese)*, 105(373):7–12.

Watanabe, T., Watanebe, B., Fujinuma, T., Osugi, N., Sawada, M., and Kamata, K. (2005b). Major factors that affect comprehensibility of shosaiyomi (explanatory expressions) used in screen readers: Consideration based on classification of shosaiyomi and kanji writing test. *The transactions of the Institute of Electronics, Information and Communication Engineers. D-I (In Japanese)*, 88(4):891–899.