The Similarity Index of Character Shape of Medicine Names Based on Character Shape Similarity (II)

Keita Nabeta[1], Akira Hatano[1], Hirotsugu Ishida[1], Masaomi Kimura[1],
Michiko Ohkura[1], and Fumito Tsuchiya[2]

[1] Shibaura Institute of Technology,
3-7-5 Toyosu, Koto-ku, Tokyo 135-8548 Japan,
[2] International University of Health and Welfare,
2600-1 Kitakanemaru, Ohtawara City, Tochigi, Japan,
{m709102,l07104,m110013,masaomi,Ohkura}@shibaura-it.ac.jp,
ftsuchiya@iuhw.ac.jp

Abstract. The similarity of drug names in Japanese such as 'アマリール' (Amaryl) and 'アルマール' (Almarl) causes confusion over drug names and can lead to medical errors. In order to prevent such errors, methods of computing their similarity have been proposed. However, there are no studies that take account of character shape similarity quantitatively. In a previous study, we calculated the character shape similarity by template matching technique and proposed a method of measuring medicine name similarity based on it. Although we obtained a high correlation coefficient between the medicine name similarity values and subjective evaluation, we observed some character pairs that are not similar. In this study, we improved the method of computing the character shape similarity based on the characteristic points of character and compared it with advanced methods.

Keywords: Medicine name similarity, Medical safety, Character shape similarity.

1 Introduction

The similarity of drug names in Japanese causes confusion over drug names and can lead to medical errors. In Japan, accidents involving confusion over the diabetes drug, 'アマリール' (Amaryl) with 'アルマール' (Almarl), which is a drug for hypertension, have resulted in patient death. In order to prevent such accidents, the Ministry of Health, Labor and Welfare has issued notices and raised awareness among medical experts. However, errors still occur.

In order to prevent such errors, it is necessary to avoid the approval of some medical names. For this purpose, many methods have been proposed for measuring the similarity of medicine names.

Tsuchiya et al. [1] proposed similarity indices for medicine names. Based on the indices, the 'Medicine similar name search engine' [2] was developed and has been operated by The Ministry of Health, Labor and Welfare. The system measures the similarity based on the head, taking account of the existence of character pairs with a

J.A. Jacko (Ed.): Human-Computer Interaction, Part I, HCII 2011, LNCS 6761, pp. 628–636, 2011.

similar shape and the position of the prolonged sound sign (dash) and the letter for a nasal sound in Japanese. However, although the method takes the similarity of character (letter) shapes into consideration, they assumed that the similarity of each character pair was given by hand [3].

In order to measure the similarity of character shape quantitatively and automatically, we applied the template matching method to katakana characters that compose Japanese medicine names [4]. We obtained a high correlation coefficient between the medicine name similarity values and subjective evaluation. However, we observed some character pairs in which the similarity values were low although they were similar. This is because the template matching technique does not take account of the connection between lines that compose katakana characters.

In this study, in order to solve the problem, we focused on the characteristic points of characters such as the edge points and intersection points. In the fields of character recognition, although methods based on characteristic points have been proposed, they also take account of other information such as surrounding characters, linguistic knowledge and the strokes of handwriting [5, 6]. However, since these are methods of recognizing characters rather than measuring the similarity of characters, we cannot utilize them. Therefore, it is necessary to develop a method of measuring the similarity of character shape based on the characteristic points of characters.

In this study, we propose a method of measuring the character shape similarity based on template matching technique and characteristic points. In order to evaluate the similarity index, we implement the character shape similarity to medicine name similarity and compare it with the subjective evaluation of pharmacists, which was obtained by an experiment.

2 Target Data

2.1 Medicine Names

In this study, we targeted the product names of ethical drugs that are included in the 'Standard Drug Master' provided by the Medical and Devices Agency (MEDIS-DC) in Japan [7].

The product names consist of brand, dosage form and ingredient amount. For instance, 'アマリール1mg錠' (Amaryl 1 mg tablet), 'アマリール' (Amaryl) is the brand part, '1 mg' is the ingredient amount and '錠' (tablet) denotes its dosage form. It is important to evaluate the similarity between brand parts, since pharmacists focus on brand parts when identifying a medicine. We therefore targeted only the brand part.

2.2 Character Type

The brand names of Japanese medicines are expressed in Hiragana, Katakana, Kanji characters, alphabets, numerical characters and other symbols. Among these character types, we concentrated on the Katakana characters since they are used to express many medicine names. This is because medicines are mainly named after foreign medicine names or active ingredient names.

Katakana characters are one component of the Japanese writing system and are often used to transcribe words from foreign languages. They are characterized by short, straight strokes and angular corners, for example ‘ア’(a), ‘イ’(i), ‘ウ’(u), ‘エ’(e), ‘オ’(o).

In this study, we used character images (height: 200px, width: 200px) generated by the Japanese character font, ‘MS Round Gothic’ (150 points) as source data.

3 Calculation of Character Shape Similarity

3.1 Template Matching [4]

The template matching technique is a general method that is used for image retrieval. By means of this algorithm, we can calculate the similarity value defined as the ratio of the number of the same colored pixels at the same location to the number of whole pixels. In this study, we digitalized font images, assigning zero to each white pixel and one to a black pixel, and calculated the similarity of each combination between all pairs of the target characters as shown in the following equation:

$$TM(a,b) = \frac{1}{mn}\sum_{j}^{n}\sum_{i}^{m}\delta(a_{ij}, b_{ij}), \tag{1}$$

where a and b are compared characters, m and n are the height and width of font images respectively and $\delta(x, y)$ is a function that returns 1 (if x is equal to y) or 0 (if x is not equal to y).

After the calculation, we realized that the obtained similarity values were high for all pairs: even the minimum similarity was 0.68. In order to redefine the similarity index so as to allow the minimum value to be zero and the maximum value to be one, we normalized them using the following equation:

$$sim_{tm}(a,b) = \frac{TM(a,b) - TM_{\min}}{TM_{\max} - TM_{\min}}, \tag{2}$$

where TM_{max} and TM_{min} denote the maximum and minimum value of similarity, respectively.

3.2 Characteristic Points

3.2.1 Extraction of Characteristic Points

Katakana characters have four characteristic points: edge points, folding points, branch points and intersection points. We extracted the edge points, the branch points and intersection points from the images of characters as shown in the following steps (Fig. 1). Firstly, we digitalized the font images. Secondly, we applied a thinning process to these images. Thirdly, we extracted characteristic points by pattern matching with three patterns as defined in Fig. 2. While for the other characteristic points, we extracted the folding points by hand since it is not easy to extract them.

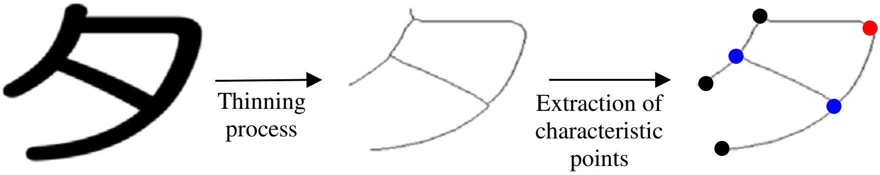
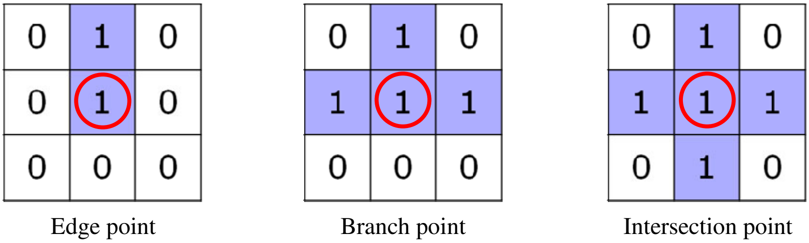


Fig. 1. Process of characteristic point extraction



0	1	0
0	1	0
0	0	0

Edge point

0	1	0
1	1	1
0	0	0

Branch point

0	1	0
1	1	1
0	1	0

Intersection point

Fig. 2. Pattern of characteristic points

3.2.2 Area Division

In order to compute the value of the similarity index using characteristic points, it is necessary to find the points corresponding to the points on another image. For this purpose, we divided the font images into different areas. If the same type points exist in the same area in each image, we can regard them as correspondence points. Figure 3 shows the distribution of characteristic points extracted from all katakana characters. The figure shows that characteristic points tend to aggregate in some parts. In this study, we considered two methods of dividing the area of font images.

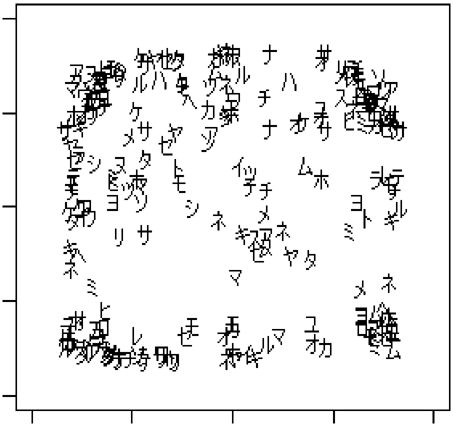


Fig. 3. Distribution of characteristic points extracted from all katakana characters

The first is a method of grid ironing the font images. The katakana characters tend to be composed of line segments extending horizontally or vertically. It is expected that dividing areas by grids takes proper account of these characters of components. However, since the most suitable grid size is not clear, we defined the size by an experiment. In addition, if there are some points near by the boundary lines of grid, they are regarded as points that belong to other areas even if they are close.

The second is a method of employing a clustering technique. The clustering technique is a method of aggregating those points into clusters that are close to each other. In this study, we utilized k-means, which is a clustering technique and aggregates the points into k clusters. By using this method, we expect that we can identify points that are close to each other to the same group. Furthermore, since the starting points and terminal points of characters tend to exist in the similar areas, it is expected to aggregate points that have the same functions. However, elongated clusters may identify separate points that should not be identified.

3.2.3 Similarity Based on Characteristic Points

If there are the same type characteristic points in the same areas, it is highly possible that they have similar characters to each other.

We calculated the absolute value of the difference between the numbers of characteristic points in an area for each type of characteristic and summed them. After computing the values in all areas, we calculated the summation of them as the distance of compared characters. Then, we divided the value by the number of all characteristic points that exist in comparing characters and subtract it from one. We defined the value as similarity based on characteristic points.

$$sim_{cp}(a,b) = 1 - \frac{\sum_D \left(\left| p_a^e - p_b^e \right| + \left| p_a^b - p_b^b \right| + \left| p_a^i - p_b^i \right| + \left| p_a^f - p_b^f \right| \right)}{\sum_D \left(p_a^e + p_b^e + p_a^b + p_b^b + p_a^i + p_b^i + p_a^f + p_b^f \right)} \tag{3}$$

where a and b are comparing characters, D are divided areas. p^e, p^b, p^i and p^f are the number of each characteristic point; edge points, branch points, intersection points and folding points.

3.3 Calculation of Character Shape Similarity

We defined the similarity index of character shape in a figure.

$$\omega_{a,b} = \alpha sim_{tm}(a,b) + (1-\alpha) sim_{cp}(a,b) \tag{4}$$

Let α be the contribution ratio between similarity values of template matching and characteristic points ($0 \le \alpha \le 1$).

4 Medicine Name Similarity

Based on the character shape similarity defined in the previous section, we computed the similarity of medicine names. In this study, we employed a method that is proposed by the advanced study [4]: Extended Letter Sequence Kernel (eLSK) and Extended Head and Tail Cosine (eHTCO).

5 Experiment

5.1 Evaluation of Character Shape Similarity

In order to evaluate our method of calculating character shape similarity, we compared it with the similarity values reported Yamade et al. [8]. In their study, they measured the similarity of katakana pairs by subjective evaluation. We use 50 katakana pairs whose similarity values are high. In order to indicate that our similarity index corresponds to a subjective view, we defined the following equation, which expresses the distance between the values of similarity, which are calculated by our method and the results obtained by Yamade's experiment.

$$D=\sum_{a,b}\left(sim_p\left(a,b\right)-sim_y\left(a,b\right)\right)^2 \tag{5}$$

where $sim_p(a, b)$ and $sim_y(a, b)$ are the similarity of the proposed method and Yamade's results between character a and b.

5.1.1 Comparison of Dividing Method

In this study, we proposed two methods of dividing the font images to compute the similarity of characters. In order to select the better method, we compared them by the evaluation index.

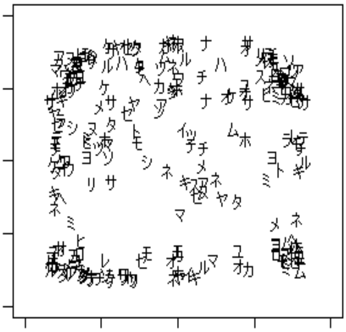


Fig. 4. Grid division (2x2)

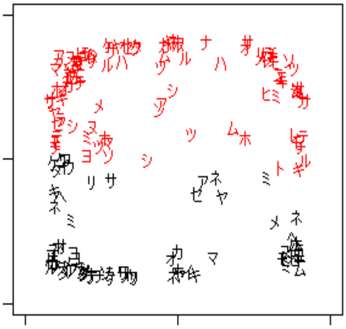


Fig. 5. Edge points (2 clusters)

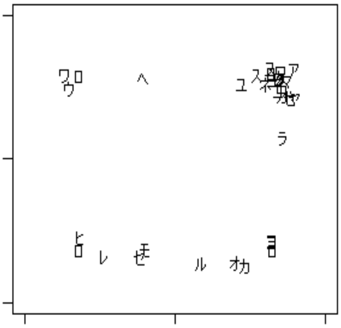


Fig. 6. Folding points (1 cluster)

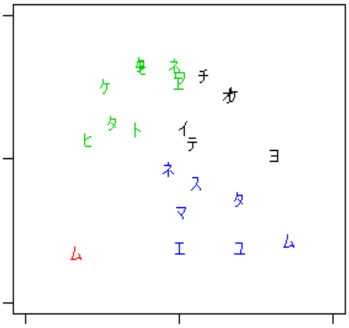


Fig. 7. Branch points (4 clusters)

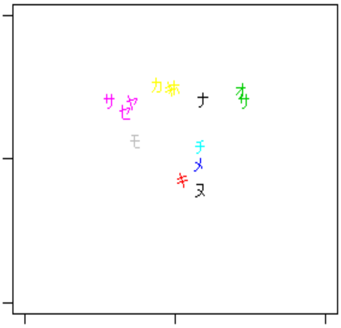


Fig. 8. Intersection points (8 clusters)

In the case of grid division, we obtained the smallest value of 1.86 when the number of horizontal division was 2 and the number of vertical division was 2 (Fig. 4). In the case of clustering division, we obtained the smallest value of 1.46 when the numbers of cluster of the edge points, the folding points, the branch points and the intersection points were 2, 1, 4 and 8 respectively (Fig. 5, 6, 7, 8). The results indicate that the latter method is better than the former method. As the reason, since most katakana characters are written from above to below or from left to right, there are edge points and folding points in the corners of font images. The superior method takes account of the features of katakana characters.

5.1.2 Comparison between Template Matching and Characteristic Point

Table 1 shows the similarity values of katakana pairs in the top 10. In the result of advanced study, we can see the dissimilar pairs whose lines match with each other such as ‘レ’ and ‘ン’. On the other hand, in the results of the proposed method, their similarity values decrease. However, there are some pairs that are not similar by the correspondence of their characteristic points such as ‘ニ’ and ‘ハ’.

Table 1. Character shape similarity in the top 10. (Yamade’s result, Template matching and Characteristic point).

A	B	Sim	A	B	Sim	A	B	Sim
シ	ツ	1.00	ク	タ	1.00	ア	ラ	1.00
ソ	ン	0.94	ソ	ツ	0.94	ア	ル	1.00
コ	ユ	0.89	コ	ニ	0.93	コ	ワ	1.00
ウ	ワ	0.88	エ	ニ	0.93	シ	ツ	1.00
シ	ソ	0.88	ク	フ	0.91	ス	マ	1.00
チ	テ	0.88	ニ	ユ	0.89	ソ	リ	1.00
シ	ン	0.86	コ	ロ	0.88	ニ	ハ	1.00
ソ	ツ	0.86	シ	ツ	0.85	ヘ	レ	1.00
ス	ヌ	0.85	エ	コ	0.84	ラ	ル	1.00
ツ	ン	0.85	シ	ン	0.84	ア	ヌ	0.88

Next, we calculated the similarity that is integrated by Equation 4. When α is 0.5, we obtained the evaluation value 0.85 by Equation 5. In addition, we can see the high correlation coefficient 0.69 between the integrated method and Yamade's result. These results indicate that it is effective to integrate the similarity indices based on the template matching and characteristic points.

5.2 Evaluation of Medicine Name Similarity

5.2.1 Method

In order to evaluate the proposed method, we compared the values computed by our method with the similarity that is evaluated by pharmacists. In the questioner, we presented the compared two medicine names on display and asked them to evaluate their similarity by values between 0 (dissimilar) and 100 (similar). In the previous study, we observed the pharmacists who answered the similarity focusing on several points such as phonological similarity. As a counter measure, we set the time (1 sec) to present the medicine names to the subjects. The number of pharmacist was 25. The names are the stem part of existing drugs. The pairs are selected so that values of similarity distribute evenly. In order to exclude the effective of length of medicine names, we selected the pairs whose medicine names are the same length. Taking account of the effect of order, we prepared 3 question patterns whose order is different from each other.

5.2.2 Results and discussion

Figure 9 shows the relation between the pharmacist evaluation and the similarity index computed by our method. These results show the correspondence between them and a height coefficient correlation of 0.87. This value is higher than the values between the pharmacist evaluation and the similarity values calculated by the method proposed in the advanced study (0.84). Furthermore, these results show the validity of the addition of similarity based on characteristic points of the existing method, which are calculated by template matching.

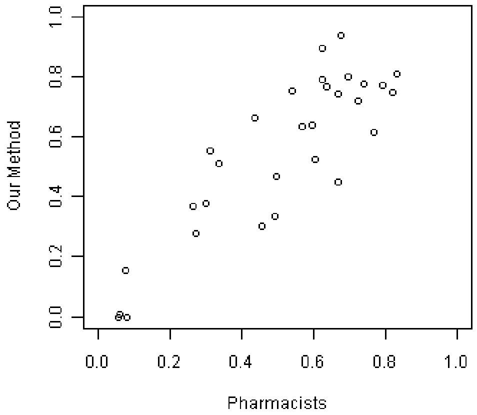


Fig. 9. Relation between pharmacist evaluation and our method

6 Conclusion

In order to ensure the medical safety of drug usage, we proposed medicine name similarity based on character shape similarity. In this study, we focused on the characteristic points of characters, which are edge points, folding points, branch points and intersection points, to redeem the method by template matching, which is proposed by the advanced study. Since it is necessary to regard points that are close to each other as corresponding points to calculate similarity using characteristic points, we proposed a method of dividing the font images into different areas. By comparing the obtained similarity values to the subjective evaluation obtained by advanced study, grid division was found to be superior to clustering division.

In order to evaluate the proposed method, we calculated the medicine name similarity based on the character shape similarity computed by the method and compared them to the similarity perceived by pharmacists. Our method was found to be superior to the advanced method, which is based on only template matching technique. Furthermore, we obtained a high correlation coefficient between our similarity and subjective evaluation of pharmacists.

In the future, it is necessary to experiment using characters that are expressed by other font types and handwriting. In addition, we should take account of the phonological similarity of characters.

References

1. Tsuchiya, F., Kawamura, N., Oh, C., Hara, A.: Standardization and similarity deliberation of Drug-names. Jpn. J. Med. Informatics 21, 59–67 (2001)
2. Japan Pharmaceutical Information Center. Medicine similar name search engine, `https://www.ruijimeisho.jp/index.aspx`
3. Ohtani, H., Takeda, M., Imada, Y., Sawada, Y.: Development of the Measures to Evaluate the Similarity of Drug Brand Names. Yakugaku Zasshi 126(5), 349–356 (2006)
4. Nabeta, K., Imai, T., Kimura, M., Ohkura, M., Tsuchiya, F.: The similarity index of medicine names to prevent confusion. In: The Pan-Pacific Conference on Ergonomics 2010 Proceedings (2010)
5. Akiyama, K., Nakagawa, M.: A Liner-Time Elastic Matching Algorithm for On-line Recognition of Handwriting Japanese Characters. The Transactions of The Institute of Electronics, Information and Communication Engineers J81-D-2(4), 651–659 (1998)
6. Nishida, S.: A Matching Algorithm of Feature Graph for Handwritten Character Recognition. IEICE Technical Report CST 108(79), 25–30 (2008)
7. The Medical Information System Development Center, `http://www.medis.or.jp/`
8. Yamade, Y., Haga, S.: Subjective evaluation of similarity of appearance of Katakana characters in drug names. Rikkyo Psychological Research 50, 79–85 (2008)