Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 35 (2014) 1519 – 1528

Procedia Computer Science

18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES2014

Similarity index for sound-alikeness of drug names with pitch accents

Tomoyuki Nagata[a], Masaomi Kimura[b,*], Fumito Tsuchiya[c]

[a]*Graduate School of Engineering and Science, Shibaura Institute of Technology, 3-7-5, Toyosu, Koto City, Tokyo, 135-8548, JAPAN*
[b]*Shibaura Institute of Technology, 3-7-5, Toyosu, Koto City, Tokyo, 135-8548, JAPAN*
[c]*International University of Health and Welfare, 2600-1 Kitakanemaru, Otawara City, Tochigi,324-8501, JAPAN*

Abstract

Drug name similarity is one of major reasons of medical accidents. In order to prevent from the accidents, one of the best ways is to avoid approving drugs that has the names similar to that of existing drugs. It is well-known that there are two kinds of drug name similarity, look-alikeness and sound-alikeness. Nabeta et. al. proposed a look-alikeness similarity index,which excludes the sound-alikeness. Though, in Japan, oral prescription is basically prohibited, emergent situation can force a doctor to prescribe orally. In such a situation, medical accidents can occur.

In this study, we proposed a sound-alikeness similarity index based on quantitative similarity of consonants. The consonant similarity was proposed based on The International Phonetic Alphabet (IPA). Overall drug name similarity is calculated based on Letter Sequence Kernel (LSK). The similarity calculation method takes account of the effect of plural pitch accents. We divided a drug name into some pieces at the position where a pitch accent changes, applied LSK to each of them, and combined them to obtain the value of the similarity index. The similarity index proposed in this study achieved relatively high correlation to the results of our experiment, $r \simeq 0.8$.

Peer-review under responsibility of KES International.

Keywords: Medical Safety, Drug names, Similarity Index

1. Introduction

In order to ensure medical safety, not only the safety of material aspect of drugs but also the safety of their usage is important. This is because, even if the drug works well to some illness, its wrong prescription to wrong patients can cause severe medical accidents.

Among the accidents that we need to prevent, the ones caused by the similarity of drug names is senseless but serious, since some patients were killed by it.

* Corresponding author. Tel.: +81-3-5859-8507 ; fax: +81-3-5859-8507.
E-mail address: masaomi@shibaura-it.ac.jp

Peer-review under responsibility of KES International.
doi:10.1016/j.procs.2014.08.235

It is well-known that there are two aspects of drug name similarity: look-alikeness and sound-alikeness. For example[1], Quelicin and Keflin are sound-alike but not look-alike, and Taxol and Taxotere are look-alike but not sound-alike. The two drug names, Almarl and Amaryl, are both look-alike and sound-alike, namely, very confusing for medical experts and sometimes cause mix-up accidents.

In Japan, the Ministry of Health, Labour and Welfare developed and operates the system that calculates the similarity indices of drug names. The similarity indices are *htco*, the ratio of common letters in head two letters and tail two letters of the names, *head*, the ratio of common letters in head three letters, *edit distance*, the number of operations to transform the name to another, and so on. The indices to measure similarity of drug names were proposed by Tsuchiya et. al.[2]

Otani et.al.[3] proposed indices named as *Htfrag* and *Vwhtfrag*, which are the extension of the Tsuchiya's *htco*.

Nabeta et.al.[4] proposed the similarity index that measure not only the similarity of letter sequences but also the shape similarity of letters therein.

The indices calculated by the system of the Ministry of Health, Labour and Welfare, the one proposed by Otani and by Nabeta are to measure look-alikeness of drug names.

The similarity index to measure sound-alikeness was proposed by Lambert et. al.[5]. They measured the similarity of consonant sequences. Based on SOUNDEX code, the consonants are categorized into six groups. Replacing consonants to group IDs, Lambert calculated the similarity based on edit distance of the ID sequences. They did not take account of vowels in similarity caluculation. This might reflect that English speakers put focus more on consonants than on vowels.

Otani also proposed an index called *Awhtfrag*, which takes account of the sound-alikeness. It measures the concidnece of either consonants or vowels in letter sequences. Phenomes are regarded to be similar if their consonants or vowels conside.

In the above studies, the quantification of consonant sound-alikeness is given in ad-hoc manner. They did not show the theoretical ground of their definition of consonant sound-alikeness. Moreover, they identified sound-alike consonants with each other and did not take account of the extent of alikeness between consonants. For most peaple, it is natural to feel that the consonants *B* and *V* are similar. However, they make the value of *Awhtfrag* lower, since they do not coincide. Though sound-alikeness between *B* and *V* and between *P* and *V* might be different, *B*, *V* and *P* are assigned the same code to by SOUNDEX, and the pair, *P* and *V*, contributes Lambert's similarity index to the same extent between *B* and *V*.

In this study, in order to take account of the extent of consonant sound-alikeness, we define objective sound-alikeness index of consonants based on the phonetic features used in the International Phonetic Alphabet (IPA). Based on the consonant sound-alikeness similarity index, we define the sound-alike similarity index of drug names. The index is designed based on Letter Sequence Kernel (LSK), which has large value if two drug names share common substrings.

The important difference of sound-alikeness and look-alikeness is whether accents affect similarity or not. Even if the spellings of two drug names are similar, the names with different accent locations are far from similar. We, therefore, need to take account of the effects of accents on drug name sound-alikeness.

We should remind that there are two kinds of accents: pitch accents and stress accents. Roughly to say, pitch accents are used in Asian languages, such as Japanese and Korean, and stress accents are used in European languages, such as English and Spanish. As our first try, we focus on Japanese drug names. It is interesting because most drug names originates in European/American drug names but their accent system is different, namely, based on Japanese-specific (pitch) accents.

あ, い, う, え, お

Moreover, Japanese language has another interesting feature that any vowel is paired with only one consonant. This is simple to discuss the effects of consonants and vowels. Japanese language has five vowels, あ (a), い (i), う (u), え (e), お (o). They are easily distinguishable and help the discussion be simple. Therefore, we focus on a discussion of consonant similarity.

As the evaluation of our similarity index, we apply it to the pairs of real drug names. We utilize Visual Analog Scale (VAS) to measure the similarity felt by subjects and compare the results with the obtained values of our index. Moreover, we compare the values of our sound-alikeness similarity index with the ones of Nabeta's similarity index, which measures look-alikeness of drug names.

2. Consonant sound-alikeness

In standard Japanese language, there are 14 consonants, which are approximately expressed as *k, s, t, n, h, m, y, l, w, g, z, d, b, p* in English alphabet.

The International Phonetic Alphabet (IPA) is an alphabetic system to represent speech sounds based on their features. As for consonants, there are two main attributes to express them: place of articulation and manner of articulation. Classification by places of articulation is given as bilabial, labiodental, dental, alveolar, postalveolar, retroflex, palatal, velar, uvular, pharyngeal and glottal, and classification by manners of articulation is given as plosive, nasal, trill, tap or flap, fricative, lateral fricative, approximant and lateral approximant. Moreover, the manners of articulation can be parametrized by presence or absence of vocal band vibration, air pathway, closing status of articulatory organ, position of velum palatinum, affrication and palatalization.

In this study, we utilize seven attributes, *places of articulation, vocal band vibration, air pathway, closing status of articulatory organ, position of velum palatinum, affrication* and *palatalization* to express consonants. In order to quantify the similarity of consonants based on these, we defined the vector whose elements correspond to 22 attribute values of the above attributes. Each element takes a binary value, namely, 1 or 0. Only one attribute value for each attribute can be 1. Let $\mathbf{c} = (c_1, c_2, \cdots, c_{22})$ and $\mathbf{c}'$ be the vectors for consonants. We designed a consonant similarity index based on vector space model, which is given by

$$sim_{cons}(\mathbf{c}, \mathbf{c}') = \frac{\mathbf{c}^T W \mathbf{c}'}{\sqrt{\mathbf{c}^T W \mathbf{c}} \sqrt{\mathbf{c}'^T W \mathbf{c}'}}, \tag{1}$$

where W is a weight matrix. The matrix W is a diagonal matrix:

$$W = diag(w_1, w_2, \cdots, w_{22}), \tag{2}$$

where w_i is a weight value.

If the element corresponds to any of the attribute values, existence of vocal band vibration, complete close of articulatory organ, blocked air pathway, low position of velum palatinum, we set $w_i = 1.2$. This is because

In contrast, the attributes, affrication or palatalization, need other consideration. For example, the difference of the consonant of キャ, *ky*, and the one of カ, *k*, is the existent of palatalization. The diffrence of the consonant of ザ, *z*, and the one of ダ,*d*, is the existent of affrication. Obviously, they are similar. This shows the existent of palatalization have less affect on similarity. Thus we set $w_i = 0.8$, if the element corresponds to affrication or palatalization.

For other elements, we set $w_i = 1$.

In this study, we introduce 27 consonant symbols: M(=my), m, B(=by), b, w, G(=gy) ,g, j, R(=ry), r, N(=ny), n, z, d, y, f(=hw), P(=py), p, K(=ky), k, H(=hy) , h, T(=ty), t, S(=sh), s, X(=ts).

Table 1 shows the values of similarity index, Eq.(1), for each pair of the 27 consonants.

Figure 1 shows the distribution of similarity index values in the descendant order. We can see the relatively long flat portion where similarity index values are around 0.78. Below this value, the change of curve slope turns to be small. This means that many pairs have similar values that are lower than 0.78. Since it is difficult to consider that most pairs of consonants are felt similar, the pairs that have the value below this threshold, 0.78, should be regarded to be not similar, namely, their similarity index values are set to be zero in similarity index calculation.

Table 2 shows the corresponding relationships between Japanese characters (J.Char.), consonants (Cons.) and vowels (Vow.).

3. Drug name sound-alikeness

3.1. Letter sequence kernel

We employed Letter sequence kernel (LSK) to define a similarity index. Tatsuno et.al.[6] proposed the utilization of LSK to quantify drug name similarity, though their target was not sound-alikeness but look-alikeness. Let us give a brief review of an extended version of this kernel[4].

Table 1. Similarity values of consonants

	M	m	B	b	w	G	g	j	R	r	N	n	z
M	1.00	-	-	-	-	-	-	-	-	-	-	-	-
m	0.92	1.00	-	-	-	-	-	-	-	-	-	-	-
B	0.81	0.73	1.00	-	-	-	-	-	-	-	-	-	-
b	0.73	0.81	0.92	1.00	-	-	-	-	-	-	-	-	-
w	0.35	0.44	0.55	0.63	1.00	-	-	-	-	-	-	-	-
G	0.68	0.59	0.87	0.78	0.55	1.00	-	-	-	-	-	-	-
g	0.59	0.68	0.78	0.87	0.63	0.92	1.00	-	-	-	-	-	-
j	0.51	0.59	0.70	0.78	0.48	0.70	0.78	1.00	-	-	-	-	-
R	0.49	0.41	0.68	0.59	0.48	0.68	0.59	0.51	1.00	-	-	-	-
r	0.41	0.49	0.59	0.68	0.57	0.59	0.68	0.59	0.92	1.00	-	-	-
N	0.87	0.78	0.68	0.59	0.28	0.68	0.59	0.51	0.62	0.54	1.00	-	-
n	0.78	0.87	0.59	0.68	0.37	0.59	0.68	0.59	0.54	0.62	0.92	1.00	-
z	0.51	0.59	0.70	0.78	0.48	0.70	0.78	0.87	0.64	0.73	0.64	0.73	1.00
d	0.59	0.68	0.78	0.87	0.57	0.78	0.87	0.78	0.73	0.81	0.73	0.81	0.92
y	0.27	0.36	0.46	0.55	0.90	0.46	0.55	0.46	0.46	0.55	0.27	0.36	0.46
f	0.22	0.30	0.41	0.49	0.44	0.27	0.36	0.27	0.27	0.36	0.08	0.17	0.27
P	0.62	0.54	0.81	0.73	0.35	0.68	0.59	0.51	0.49	0.41	0.49	0.41	0.51
p	0.54	0.62	0.73	0.81	0.44	0.59	0.68	0.59	0.41	0.49	0.41	0.49	0.59
K	0.49	0.41	0.68	0.59	0.35	0.81	0.73	0.51	0.49	0.41	0.49	0.41	0.51
k	0.41	0.49	0.59	0.68	0.44	0.73	0.81	0.59	0.41	0.49	0.41	0.49	0.59
h	0.08	0.17	0.27	0.36	0.37	0.27	0.36	0.27	0.27	0.36	0.08	0.17	0.27
T	0.32	0.41	0.51	0.59	0.28	0.51	0.59	0.68	0.32	0.41	0.32	0.41	0.68
S	0.17	0.08	0.36	0.27	0.42	0.36	0.27	0.19	0.49	0.41	0.30	0.22	0.32
s	0.08	0.17	0.27	0.36	0.51	0.27	0.36	0.27	0.41	0.49	0.22	0.30	0.41
X	0.32	0.41	0.51	0.59	0.28	0.51	0.59	0.68	0.45	0.54	0.45	0.54	0.81
t	0.41	0.49	0.59	0.68	0.37	0.59	0.68	0.59	0.54	0.62	0.54	0.62	0.73
H	0.08	0.17	0.27	0.36	0.51	0.27	0.36	0.27	0.27	0.36	0.08	0.17	0.27

	d	y	f	P	p	K	k	h	T	S	s	X	t	H
d	1.00	-	-	-	-	-	-	-	-	-	-	-	-	-
y	0.55	1.00	-	-	-	-	-	-	-	-	-	-	-	-
f	0.36	0.36	1.00	-	-	-	-	-	-	-	-	-	-	-
P	0.59	0.27	0.59	1.00	-	-	-	-	-	-	-	-	-	-
p	0.68	0.36	0.68	0.92	1.00	-	-	-	-	-	-	-	-	-
K	0.59	0.27	0.46	0.87	0.78	1.00	-	-	-	-	-	-	-	-
k	0.68	0.36	0.55	0.78	0.87	0.92	1.00	-	-	-	-	-	-	-
h	0.36	0.36	0.87	0.46	0.55	0.46	0.55	1.00	-	-	-	-	-	-
T	0.59	0.27	0.46	0.70	0.78	0.70	0.78	0.46	1.00	-	-	-	-	-
S	0.41	0.41	0.65	0.55	0.46	0.55	0.46	0.65	0.38	1.00	-	-	-	-
s	0.49	0.49	0.74	0.46	0.55	0.46	0.55	0.74	0.46	0.92	1.00	-	-	-
X	0.73	0.27	0.46	0.70	0.78	0.70	0.78	0.46	0.87	0.51	0.59	1.00	-	-
t	0.81	0.36	0.55	0.78	0.87	0.78	0.87	0.55	0.78	0.59	0.68	0.92	1.00	-
H	0.36	0.62	0.74	0.46	0.55	0.46	0.55	0.74	0.46	0.78	0.87	0.46	0.55	1.00

Let us define a function that measures the contribution of the substring u in the string s:

$$\phi_u(s) = \sum_{s[i_1,i_2,\cdots,i_n]} \left(\prod_j \omega_{u[i],s[i_j]} \right) \lambda^{(i_n - i_i + 1) - n} \tag{3}$$

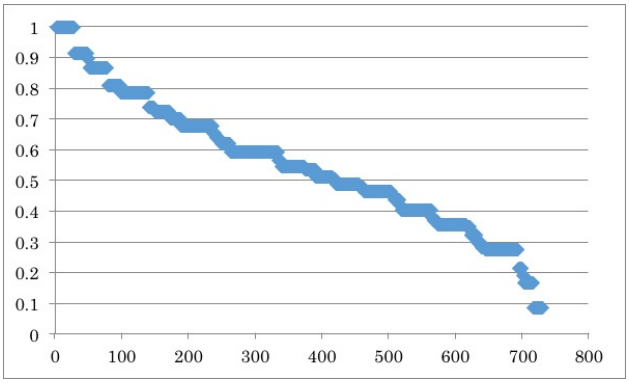

Fig. 1. **The similarity index values in the descendant order versus their ranks. The horizontal axis denotes rank and the vertical axis denotes similarity index value.**

where n is string length of u, $s[i_1, i_2, \cdots, i_n]$ denotes a substring of s whose length is equal to n, λ is a constant less than 1 and $u[i]$ is the i^{th} letter of u. This equation takes account of letter similarity contained in both u and s by introducing $\omega_{a,b}$, which denotes the letter similarity index between two letters a and b. If the positions of the first letter and the last letter of the substring $s[i_1, i_2, \cdots, i_n]$ are farther than n, other characters in s exist between its characters. This makes it and u less similar. In such a case, the factor $\lambda^{(i_n - i_i + 1) - n}$ makes its contribution small in $\phi_u(s)$.

Regarding this is an element of the vector representing s, an inner product of the vectors for strings s and t can be given as:

$$K(s,t) = \sum_{u \in s \wedge u \in t} \phi_u(s)\phi_u(t). \tag{4}$$

This $K(s,t)$ is LSK.

Utilizing this, we can define a similarity index as followings:

$$sim_{LSK}(s,t) = \frac{K(s,t)}{\sqrt{K(s,s)}\sqrt{K(t,t)}}. \tag{5}$$

3.2. Accents

In Japanese language, there are the words that have the same spelling but different accents. An example is the word,“はし” (its sound is expressed as “haSi” in our notation), which has different meaning depending on its pitch series. If the first letter “は”(ha) is in high pitch and next letter “し”(Si) in lower, it means a bridge or chop sticks. If the first letter low and next high, it means a corner or an edge. This suggests that even if the spelling is same, different pitches make us feel different.

In this study, we assume that words sound similar if each of their segments divided by pitch changes sounds similar. The good example is how we pronounce a composite word. Such a word consists of plural words, which are usually segmented by pitch changes. Even if one of corresponding segments sounds different, the whole word should sound different.

Our method, therefore, segments drug names by their pitch changes, calculates similarity index values for the corresponding segments and combine the values to obtain the drug name similarity value.

3.3. Drug name sound-alikeness similarity index

Based on the above discussion, we propose the drug name sound-alikeness similarity index.

We illustrate each step based on Japanese drug names “アルマール (Almarl)” , which is expressed as “Aa ru ma 1a ru” in our convention and “アマリール (Amaryl)” expressed as “Aa ma ri 2i ru”.

Table 2. The correspondence between Japanese character, consonants and vowels.

J.Char.	Cons.	Vow.	J.Char.	Cons.	Vow.	J.Char.	Cons.	Vow.	J.Char.	Cons.	Vow.
ファ	f	a	ズ	z	u	ム	m	u	ド	d	o
ア	A	a	ゼ	z	e	メ	m	e	ヂャ	D	a
イ	I	i	ゾ	z	o	モ	m	o	ヂ	D	i
ウ	U	u	ジャ	j	a	ミャ	M	a	ヂュ	D	u
エ	E	e	ジ	j	i	ミ	M	i	ヂェ	D	e
オ	O	o	ジュ	j	u	ミュ	M	u	ヂョ	D	o
カ	k	a	ジェ	j	e	ミェ	M	e	ヅ	*	u
ク	k	u	ジョ	j	o	ミョ	M	o	ハ	h	a
ケ	k	e	ナ	n	a	ラ	r	a	ヘ	h	e
コ	k	o	ヌ	n	u	ル	r	u	ホ	h	o
キャ	K	a	ネ	n	e	レ	r	e	ヒャ	H	a
キ	K	i	ノ	n	o	ロ	r	o	ヒ	H	i
キュ	K	u	ニャ	N	a	リャ	R	a	ヒュ	H	u
キェ	K	e	ニ	N	i	リ	R	i	ヒェ	H	e
キョ	K	o	ニュ	N	u	リュ	R	u	ヒョ	H	o
ガ	g	a	ニェ	N	e	リェ	R	e	フィ	f	i
グ	g	u	ニョ	N	o	リョ	R	o	フ	f	u
ゲ	g	e	バ	b	a	タ	t	a	フェ	f	e
ゴ	g	o	ブ	b	u	ティ	t	i	フォ	f	o
ギャ	G	a	ベ	b	e	トゥ	t	u	ヤ	y	a
ギ	G	i	ボ	b	o	テ	t	e	ユ	y	u
ギュ	G	u	ビャ	B	a	ト	t	o	イェ	y	e
ギェ	G	e	ビ	B	i	チャ	T	a	ヨ	y	o
ギョ	G	o	ビュ	B	u	チ	T	i	ワ	w	a
サ	s	a	ビェ	B	e	チュ	T	u	ウィ	w	i
スィ	s	i	ビョ	B	o	チェ	T	e	ウェ	w	e
ス	s	u	パ	p	a	チョ	T	o	ウォ	w	o
セ	s	e	プ	p	u	ツァ	X	a	ッ	X	x
ソ	s	o	ペ	p	e	ツィ	X	i	ン	/	n
シャ	S	a	ポ	p	o	ツ	X	u	ヴァ	b	a
シ	S	i	ピャ	P	a	ツェ	X	e	ヴ	b	u
シュ	S	u	ピ	P	i	ツォ	X	o	ヴェ	b	e
シェ	S	e	ピュ	P	u	ダ	d	a	ヴォ	b	o
ショ	S	o	ピェ	P	e	ディ	d	i	ヴィ	B	i
ザ	z	a	ピョ	P	o	ドゥ	d	u			
ズィ	z	i	マ	m	a	デ	d	e			

The part “1a” and “2i” appearing here denotes prolonged sound. Since prolonged sound in アルマール sits just after the sound “マ”, whose vowel is “a”, we assign the virtual consonant “1” to this prolonged sound. We distinguish the difference of prolonged sounds based on the sound just before it. Thus the prolonged sound in アマリール is denoted as “2i”, since it sits after the vowel “i”.

The similarity index calucuration is realized as the following steps.

1. Divide each name into its consonant part and vowel part. The アルマール is decomposed to the consonant part “Arm1r” and the vowel part “auaau”. The アマリール is decomposed to “Amr2r” and “aaiiu”.

2. Segment the consonant parts and the vowel parts at the pitch changing position. As for アルマール, the pitch changes after the first letter "ア" and the letter "マ". Therefore, its consonant part is segmented into the three parts, "A", "rm" and "1r", and its vowel part is "a", "ua" and "au". As for アマリール, we obtain the consonant segments, "A", "mr" and "2r" and the vowel segments, "a", "ai" and "iu".
3. Calculate LSK similarity values for each corresponding pairs:

   - $c_1 = sim_{LSK}(A, A) = 1.00$,
   - $c_2 = sim_{LSK}(rm, mr) = 0.784$,
   - $c_3 = sim_{LSK}(1r, 2r) = 0.654$,
   - $v_1 = sim_{LSK}(a, a) = 1.00$,
   - $v_2 = sim_{LSK}(ua, ai) = 0.333$,
   - $v_3 = sim_{LSK}(au, iu) = 0.333$.

4. Calculate "segment similarity" by taking their products,

   - $c_1 \cdot v_1 = 1.00$,
   - $c_2 \cdot v_2 = 0.261$,
   - $c_3 \cdot v_3 = 0.218$.

5. Sort the segment similarity values and multiply $\mu = 1.5$ to the largest value and $\mu = 0.5$ to the smallest value.
6. Take an arithmetic mean to the weighted segment similarity values, $\frac{1.5\times1.00+1.0\times0.261+0.5\times0.218}{1.5+1+0.5} = 0.623$.

The formal expression of our index is given as:

$$sim_{accent}(s,t) = \frac{\sum_i \left(\mu_i \cdot sim_{LSK}(s_c^{(i)}, t_c^{(i)}) \cdot sim_{LSK}(s_v^{(i)}, t_v^{(i)})\right)}{\sum_i \mu_i}, \tag{6}$$

where $s_c^{(i)}$ is the consonant part of the i^{th} segment of Drug name s, and $s_v^{(i)}$ is its vowel part.

4. Experiments

4.1. Comparison with VAS values

We conducted an experiment to compare how similar subjects feel with our similarity index. We utilized Visual Analog Scale (VAS) method to measure subjects' feeling of similarity. We presented 40 pairs of drug name stems chosen out of 7,727 Japanese drug stems to subjects, and asked them to answer how similar they were in the scale between 0 (completely different) and 100 (same).

In order to prevent the distribution of similarity values from distorting, we chose the pairs of drug names so that they contain the pairs which are different combinations of the segments that have high pitch. The difference of and the pairs whose names are different up to 4 syllables.

The subjects were 22 university school students who majored in computer engineering. This is because the subjects who do not have medical knowledge is suitable to measure the sound-alikeness of drug names without the bias originating in medical pre-knowledge.

As a result, the mean VAS values and the proposed similarity index values have relatively high correlation, whose correlation coefficient is 0.719. This shows that, to some extent, our index can predict how similar subjects feel.

We also found that the pairs with the same head letter tends to give large VAS values compared to our proposed index. Let $head_2(s,t)$ be the coincident number of head two letters of String s and t. After some experiments, we found that inclusion of the contribution of $head_2(s,t)$,

$$sim_{sound}(s,t) = 0.2\frac{head_2(s,t)}{2} + 0.8 sim_{accent}(s,t), \tag{7}$$

improves the correlation with mean VAS values. The correlation coefficient for the mean VAS values and this index values is 0.805 (Figure 2).

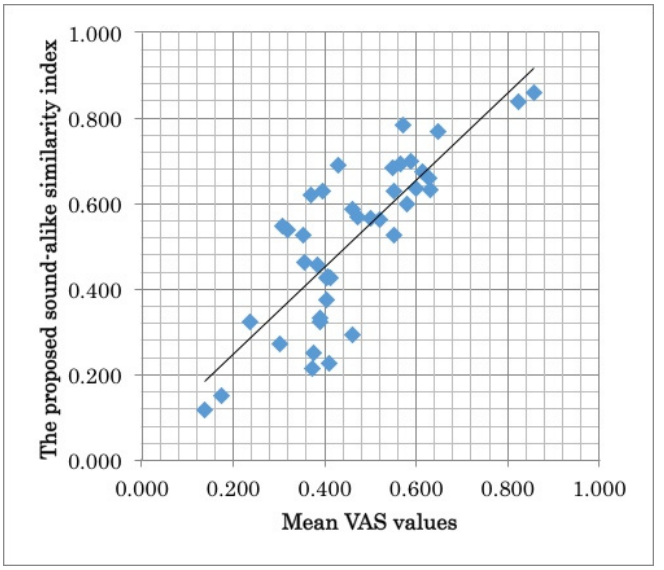

Fig. 2. **The mean VAS values VS Sound-alikeness similarity index**

4.2. Comparison with look-alikeness similarity index

We calculated our sound-alikeness similarity index with Nabeta's look-alikeness similarity index and compared them to see how different characters appear in those indices. The target data is a set of drug names any of whose edit distance are less than 3.

Figure 3 shows the results. We can easily see that there are some pairs both look-alike and sound-alike. Some other pairs are either look-alike or sound-alike.

Table 3 shows typical types of alikeness. Type A groups the pairs both look-alike and sound-alike. They can confuse medical experts when they are in emergent condition. In order to prevent such confusion, they need some countermeasure, such as addition of information such as dosage forms or standard units to make the difference clear. Type B groups the pairs sound-alike not look-alike. They need caution if a medical expert tells their names to other one. Type C groups the pairs look-alike not sound-alike. They can confuse a medical expert when he read their names written on a drug package in prescriptions.

The drug names that are sound-alike or look-alike to other names should be changed or, at least, needs caution to prevent such confusion.

Table 3. Look-alikeness vs Sound-alikeness.

Type	Drug name 1	Drug name 2	Look-alikeness	Sound-alikeness
A	アスプール (ASTHPUL)	アスクール (ASCOOL)	0.821	0.999
A	ベノジール (BENOZIL)	ベノキシール (BENOXIL)	0.831	0.827
B	ボグリース (VOGLISE)	ボグシール (VOGSEAL)	0.495	0.959
B	アルカドール (ALCADOL)	アルナゾール (ALNAZOL)	0.674	0.959
C	バルネチール (BARNETIL)	パルチデール (PALTIDEL)	0.770	0.450
C	バイコール (BAYCOL)	ハイコバール (HYCOBAL)	0.851	0.435

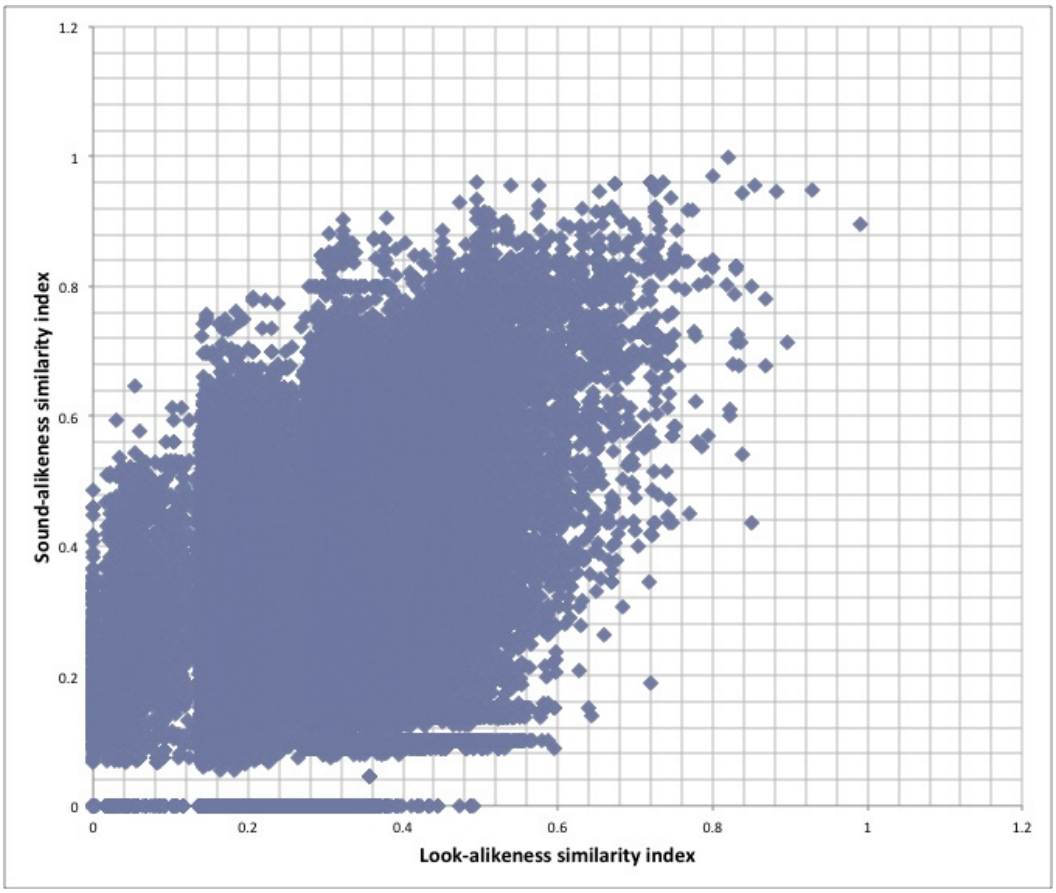

Fig. 3. **Sound-alikeness similarity index VS look-alikeness similarity index**

5. Conclusions

In this paper, we discussed the sound-alikeness of drug names and proposed the sound-alikeness similarity index taking account of quantified consonant sound-alikeness and pitch accent location.

In order to define consonant sound-alikeness index, we employed the idea based on The International Phonetic Alphabet (IPA). Overall drug name similarity is calculated based on Letter Sequence Kernel (LSK). Our method segments the consonant parts and the vowel parts at the pitch changing position, applies LSK calculation to the segments. Their weighted mean is defined as the index,sim_{accent}. Taking account of head letters' effect, we add this sim_{accent} and $head_2/2$, the ratio of head two letters' coincidence, in the ratio, 0.8 : 0.2.

This achieved high correlation to the similarity that is felt by subjects in our experiments. Concretely, we obtained high correlation coefficient of +0.805 between the mean VAS values answered by subjects in our experiments and our proposed sound-alikeness similarity index values.

We also found the pairs of drug names which are either or both look-alike and/or sound-alike based on our proposed index and Nabeta's look-alikeness similarity index.

References

1. WHO Collaborating Centre for Patient Safety Solutions. Look-alike, sound-alike medication name, http://www.who.int/patientsafety/solutions/patientsafety/PS-Solution1.pdf (accessed on May 13th 2014)

2. Tsuchiya F, et.al. Standardization and similarity deliberation of Drug-names, *Japan journal of medical informatics* 21(1), 59-67, 2001.
3. Ohtani H. et.al. Development of the Measures to Evaluate the Similarity of Drug Brand Names,*Journal of the Pharmaceutical Society of Japan* 126(5), 2006, 349-356.
4. Nabeta, K., et.al.: A Proposal of Method to Calculate Similarity of Medicine Brand Names Based on Character Resemblance,*Transactions of Japan Society of Kansei Engineering* 10(2), 2011, 287-294.
5. Lambert, B. L., Lin S-J., Gandhi, S. K.,Chang K-Y: Similarity as a risk factor in drug name confusion errors: The look-alike (orthographic) and sound-alike (phonological) model, *Medical Care*, Vol.37, 1999, pp.1214-1225.
6. Tatsuno, K., et.al.: The Study of Similarity Index of the Name of Drugs, *IEICE technical report* 106(431), 2006, 1-4.